Journal of Machine Learning Research 26 (2025) 1-70 Submitted 11/22; Revised 7/24; Published 2/25

Unbalanced Kantorovich-Rubinstein distance, plan, and barycenter on finite spaces: A statistical perspective

Shayan Hundrieser* S.HUNDRIESER@MATH.UNI-GOETTINGEN.DE
Florian Heinemann* FLORIAN.HEINEMANN@UNI-GOETTINGEN.DE
Marcel Klatt MKLATT@MATHEMATIK.UNI-GOETTINGEN.DE
Marina Struleva MARINA.STRULEVA@UNI-GOETTINGEN.DE
Axel Munk† MUNK@MATH.UNI-GOETTINGEN.DE
Institute for Mathematical Stochastics
University of Göttingen
Goldschmidtstraße 7, 37077 Göttingen

Editor: Quentin Berthet

Abstract

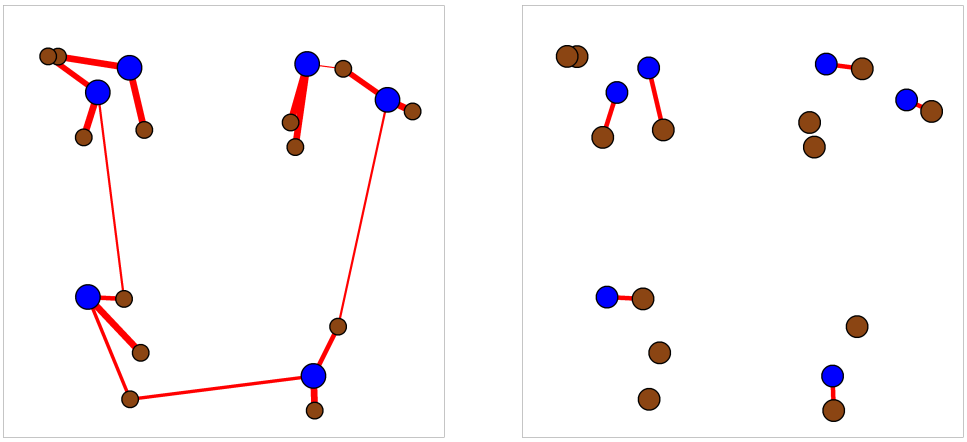
We analyze statistical properties of plug-in estimators for unbalanced optimal transport quantities between finitely supported measures in different prototypical sampling models. Specifically, our main results provide non-asymptotic bounds on the expected error of empirical Kantorovich-Rubinstein (KR) distance, plans, and barycenters for mass penalty parameter $C > 0$. The impact of the mass penalty parameter C is studied in detail. Based on this analysis, we mathematically justify randomized computational schemes for KR quantities which can be used for fast approximate computations in combination with any exact solver. Using synthetic and real datasets, we empirically analyze the behavior of the expected errors in simulation studies and illustrate the validity of our theoretical bounds.

Keywords: Unbalanced optimal transport, Barycenters, Statistical deviation bounds, (p, C)-Kantorovich-Rubinstein distance, Tree-Approximation, Resampling

1. Introduction

Optimal transport (OT) (for a detailed mathematical discussion see e.g. Villani, 2008; Santambrogio, 2015) has been a focus of attention in various research fields for a long time. More recently, its powerful geometric features promoted by improved computational tools (see e.g. Chizat et al., 2018a; Peyré and Cuturi, 2019; Guo et al., 2020; Lin et al., 2020) have turned OT into a promising new tool for modern data analysis with applications in machine learning (Frogner et al., 2015; Arjovsky et al., 2017; Schmitz et al., 2018; Yang et al., 2018; Vacher et al., 2021), computer vision (Baumgartner et al., 2018; Kolkin et al., 2019), computational biology (Evans and Matsen, 2012; Gellert et al., 2019; Klatt et al.,

∗. These authors contributed equally
†. Additionally: Max Planck Institute for Multidisciplinary Science, Am Faßberg 11, 37077 Göttingen and University Medical Center Göttingen, Cluster of Excellence 2067 Multiscale Bioimaging - From molecular machines to networks of excitable cells



(a) (b)

Figure 1: Transport between two measures (blue and brown) with their support points located in $[0,1]^2$. The respective transport plans between them are displayed by red lines where the thickness of a line is proportional to the transported mass. **(a)** The measures have been normalized to probability measures (the blue points have mass $1/6$, the brown ones have mass $1/13$). **(b)** All points in the unnormalized measures have mass 1, therefore the UOT plan for the $(2,2)$-KRD yields a one-to-one matching between a sub-collection of points.

2020; Tameling et al., 2021; Bunne et al., 2023), image processing (Pitié et al., 2007; Bonneel et al., 2016; Tartavel et al., 2016) and statistical inference (Sommerfeld and Munk, 2018; Lee and Raginsky, 2018; Mena and Niles-Weed, 2019; Panaretos and Zemel, 2019; Hallin et al., 2021; Hallin, 2022), among others.

The wide range of applications also surfaced limitations of classical OT. In particular, the assumption of equal total mass intensity of the measures is often inappropriate, e.g., in the context of image processing and retrieval (Rubner et al., 1998; Pele and Werman, 2008, 2009; Rabin and Papadakis, 2015), for takings means of neuroimaging data (Gramfort et al., 2015), when comparing radiation patterns of collider events (Komiske et al., 2019; Manole et al., 2024b), in recovery of developmental trajectories of embryonic stem cells (Schiebinger et al., 2019; Ventre et al., 2023), and in multi-color super-resolution colocalization analysis (Naas et al., 2024). If standard OT methodology had been used in these contexts, all measures would have needed to be normalized in order to overcome the issue of different total mass. However, this preprocessing step would structurally change the problem and directly impact the data analysis, in particular the underlying transport plan. For example, when matching point clouds of different sizes the resulting plan distributes mass among several points, whereas often it is desired to match points one-to-one which is favorable in many applications (see Figure 1). Attempts to circumvent this issue have led to a range of *unbalanced optimal transport* (UOT) proposals (see above applications and also Figalli, 2010; Liero et al., 2018; Chizat et al., 2018b; Balaji et al., 2020; Chapel et al., 2020; Le et al., 2021, 2022; Mukherjee et al., 2021; Heinemann et al., 2023). These formulations

extend OT concepts to general positive measures by either fixing the total amount of mass to be transported in advance or by penalizing the hard marginal constraints inherent in OT. These approaches also give rise to barycenters, generalizing the popular notion of *OT barycenters* (Agueh and Carlier, 2011) to measures of unequal mass (Gramfort et al., 2015; Chizat et al., 2018a; Friesecke et al., 2021; Heinemann et al., 2023).

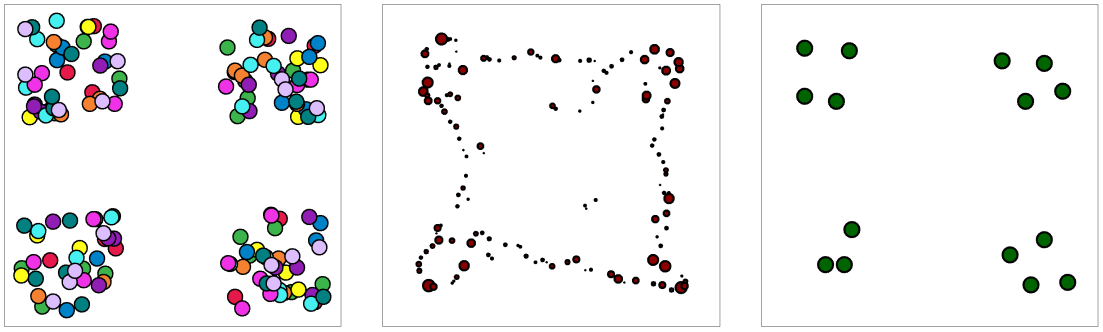
The UOT formulation considered here is the (p, C)-Kantorovich-Rubinstein distance (KRD) (see Section 1.1) whose structural properties and related (p, C)-barycenter have recently been studied in detail (Heinemann et al., 2023). The $(1, 1)$-KRD essentially corresponds to the notion of extended Kantorovich norms (Kantorovich and Rubinstein, 1958; Hanin, 1992) considered in the context of Lipschitz spaces and signed measures. In this context, Guittet (2002) first introduced a discrete formulation of the problem, where he established a Linear Program (LP) formulation which carries over to the general (p, C)-KRD. For illustration, a comparison between the (p, C)-barycenter and the p-Wasserstein barycenter in a simple example is displayed in Figure 2. From a data analysis point of view we find it particularly appealing that for the (p, C)-KRD there is a clear geometrical connection between its penalty C and the structural properties of the corresponding UOT plans and (p, C)-barycenter. More precisely, it is shown in (Heinemann et al., 2023, Lemma 2.1) that C controls the largest scale at which mass transport is possible in an optimal plan. This interpretation of the (p, C)-KRD allows designing it to respect different structural properties of the data and thus makes it a prime candidate for statistical tasks in data analysis. In particular, this emphasizes the robustness of the (p, C)-KRD to spatial outliers, i.e., data points which are located far in the tails. Furthermore, each support point of any (p, C)-barycenter is contained in a finite set characterized by the value of C. This allows to adapt OT solvers for the unbalanced problems.

Due to the unbalanced nature of the problem, however, the task of sampling from the underlying measures requires alternative sampling schemes and different statistical modelling. While for OT between probability measures there is a canonical sampling model by i.i.d. replications from the measures, this fails for UOT, since the considered measures are not necessarily probability measures. In this work, we address this issue and suggest a framework underpinning UOT based statistical data analysis. To this end, we analyze the (p, C)-KRD, the corresponding UOT plan and its barycenter in three specific statistical models motivated by applications in randomized algorithms and microscopy tasks. Notably, these models also provide a framework which potentially allows treating the alternative UOT models mentioned above. Throughout, we focus on finite discrete domains, as it allows a complete analysis while imposing minimal assumptions on the ground space.

1.1 Kantorovich-Rubinstein Quantities: Distance, Plan, and Barycenter

Let $(\mathcal{X}, d)$ be a finite metric space with finite cardinality $M \coloneqq |\mathcal{X}|$ and denote by $\mathcal{M}_+(\mathcal{X}) \coloneqq \left\{\mu \in \mathbb{R}^M \mid \mu(x) \geq 0 \ \forall x \in \mathcal{X}\right\}$ the set of non-negative measures[1] on $\mathcal{X}$. The total mass of a measure $\mu \in \mathcal{M}_+(\mathcal{X})$ is defined as $\mathbb{M}(\mu) \coloneqq \sum_{x \in \mathcal{X}} \mu(x)$ and the subset $\mathcal{P}(\mathcal{X}) \subset \mathcal{M}_+(\mathcal{X})$

1. A non-negative measure on a finite space $\mathcal{X}$ is uniquely characterized by the values it assigns to each singleton $\{x\}$. To ease notation we write $\mu(x)$ instead of $\mu(\{x\})$. The corresponding σ-field is always to be understood as the power set of $\mathcal{X}$.



(a) (b) (c)

Figure 2: **(a)** $J = 10$ measures with mass 1 at each support point and different total mass intensities (each color corresponds to a different measure) superimposed on top of each other. **(b)** The OT barycenter (for squared Euclidean cost) of the normalized measures. **(c)** The $(2, 0.3)$-barycenter of the unnormalized measures (see (3) for a rigorous definition).

of measures with total mass one is the set of probability measures. If $\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{X})$ is a measure on the product space $\mathcal{X} \times \mathcal{X}$ its marginals are defined as $\pi(x, \mathcal{X}) := \sum_{x'} \pi(x, x')$ and $\pi(\mathcal{X}, x') := \sum_{x' \in \mathcal{X}} \pi(x, x')$, respectively. For two measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ define the set of *non-negative sub-couplings* as

$$\Pi_{\leq}(\mu, \nu) := \{\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{X}) \,|\, \pi(x, \mathcal{X}) \leq \mu(x),\, \pi(\mathcal{X}, x') \leq \nu(x') \,\forall\, x, x' \in \mathcal{X}\}. \tag{1}$$

Following Heinemann et al. (2023), for $p \geq 1$ and a parameter $C > 0$, the $(p, C)-$*Kantorovich-Rubinstein distance* (KRD) between two measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ is defined as

$$\mathrm{KR}_{p,C}(\mu, \nu) := \left(\min_{\pi \in \Pi_{\leq}(\mu,\nu)} \sum_{x,x' \in \mathcal{X}} d^p(x, x')\pi(x, x') + C^p \left(\frac{\mathbb{M}(\mu) + \mathbb{M}(\nu)}{2} - \mathbb{M}(\pi) \right) \right)^{\frac{1}{p}}. \tag{2}$$

For $p \geq 1$, it defines a distance on $\mathcal{M}_+(\mathcal{X})$ robust to spatial outliers[2] and naturally extends the well-known *p-th order OT distance* defined for measures of equal total mass (Heinemann et al., 2023). Notably, by compactness of $\Pi_{\leq}(\mu, \nu)$ and continuity of the objective, an optimizer $\pi \in \Pi_{\leq}(\mu, \nu)$ of (2) always exists and is called *unbalanced optimal transport plan* (UOT plan).

The (p, C)-KRD also allows defining a notion of a barycenter for a collection of measures with (potentially) different total masses. Assume $(\mathcal{X}, d)$ to be embedded in some connected ambient space[3] $(\mathcal{Y}, d)$, e.g., a Euclidean space, and define the (p, C)-*Fréchet functional*

$$F_{p,C}(\mu) = \frac{1}{J} \sum_{i=1}^{J} \mathrm{KR}^p_{p,C}(\mu^i, \mu). \tag{3}$$

2. As a toy example, let $\mu_{x,\alpha} := \alpha\delta_0 + (1-\alpha)\delta_x$ and $\nu_\alpha := \alpha\delta_0 + (1-\alpha)\delta_1$ for $x \geq 0$ and $\alpha \in [0, 1]$. Then, $\mathrm{KRD}_{1,C}(\mu_{x,\alpha}, \nu_\alpha) = (1-\alpha)\min(|x-1|, C)$ and $W_1(\mu_{x,\alpha}, \nu_\alpha) = (1-\alpha)|x-1|$, asserting for $x \to \infty$ and $\alpha \nearrow 1$ that $\mathrm{KRD}_{1,C}(\mu_{x,\alpha}, \nu_\alpha) \to 0$; but if $\alpha = 1 - \min(1, x^{-1+\varepsilon})$ for $\varepsilon > 0$, then $W_1(\mu_{x,\alpha}, \nu_\alpha) \to \infty$.
3. We assume the metric on $\mathcal{X} \subset \mathcal{Y}$ to be the metric of $\mathcal{Y}$ restricted to $\mathcal{X}$.

Any minimizer of this functional in $\mathcal{M}_+(\mathcal{Y})$ is said to be a (p, C)-*Kantorovich-Rubinstein barycenter* of $\mu_1, \ldots, \mu_J$ or (p, C)-*barycenter* for short[4]. The objective functional $F_{p,C}$ is referred to as (unbalanced) (p, C)-*Fréchet functional* and the so-called *Borel barycenter application* is defined as $T^{L,p}(x_1, \ldots, x_L) \in \operatorname{argmin}_{y \in \mathcal{Y}} \sum_{i=1}^{L} d^p(x_i, y)$. Define the *full centroid set*[5] of the measures

$$\begin{aligned} \mathcal{C}_{\mathrm{KR}}(J, p) = \Big\{ y \in \mathcal{Y} \mid \exists L \geq \lceil J/2 \rceil,\ \exists (i_1, \ldots, i_L) \subset \{1, \ldots, J\}, \\ x_1, \ldots, x_L :\ x_l \in \operatorname{supp}(\mu_{i_l}) \\ \forall l = 1, \ldots, L :\ y = T^{L,p}(x_1, \ldots, x_L) \Big\}, \end{aligned} \tag{4}$$

and based on it the *restricted centroid set*

$$\begin{aligned} \mathcal{C}_{\mathrm{KR}}(J, p, C) = \Big\{ y = T^{L,p}(x_1, \ldots, x_L) \in \mathcal{C}_{\mathrm{KR}}(J, p) \mid \forall 1 \leq l \leq L : \\ d^p(x_l, y) \leq C^p;\ \sum_{i=1}^{L} d^p(x_l, y) \leq \frac{C^p (2L - J)}{2} \Big\}. \end{aligned} \tag{5}$$

According to (Heinemann et al., 2023, Theorem 2.5), any (p, C)-barycenter is finitely supported, and its support is included in the restricted centroid set $\mathcal{C}_{KR}(J, p, C)$. This is critical, as it allows us to restrict theoretical analysis as well as computational methods to the scenario all measures involved have finite support. Moreover, it permits an additional degree of interpretability of the mass penalization parameter C which might be of interest for specific applications.

1.2 On Plug-in Estimators for Optimal Transport

In practice, one often does not have access to the population measures $\mu, \nu, \mu^1, \ldots, \mu^J$, respectively, and they need to be estimated from data. For probability measures the most common statistical model assumes access to i.i.d. data $X_1, \ldots, X_N \sim \mu$ (and similar for $\nu, \mu^1, \ldots, \mu^J$). A commonly used estimator is the empirical measure $\widehat{\mu}_N = (1/N) \sum_{k=1}^{N} \delta_{X_k}$, where δ_X denotes the (random) point measure at location X. In light of the ongoing field of statistical OT, there are various topics of interest which have been extensively analyzed.

First, the metric property of the p-Wasserstein distance W_p (see, e.g., Villani (2008)) allows using it to evaluate the statistical accuracy of the estimator $\widehat{\mu}_N$ in estimating μ (Dudley, 1969; Fournier and Guillin, 2015; Weed and Bach, 2019). On the other hand, it is of similar interest to investigate the behavior of $W_p(\widehat{\mu}_N, \widehat{\nu}_N)$ as an estimator for the true functional $W_p(\mu, \nu)$, which in general yields statistically efficient estimators (Sommerfeld et al., 2019; Staudt and Hundrieser, 2023; Hundrieser et al., 2024c; Manole and Niles-Weed, 2024), although for continuous settings under appropriate smoothness assumptions improvements

4. For the sake of readability, the weights in this definition are fixed to $1/J$. Adaptation of all results to arbitrary positive weights $\lambda_1, \ldots, \lambda_J$ summing to one is straightforward.
5. There are scenarios where multiple sets fulfil the definition of the centroid set, since there might be multiple points that minimize the barycentric application. In this case, a fixed representative is chosen and there still exists a choice of centroid set which contains the support of the (p, C)-barycenter.

are possible (Niles-Weed and Berthet, 2022). In addition to that, considerable interest has been put in deriving distributional limits for the OT plan (Klatt et al., 2022; Liu et al., 2023) for the discrete setting as well as estimating the OT map in the continuous setting (Hütter and Rigollet, 2021; Deb et al., 2021b; Manole et al., 2024a) under smoothness assumptions and in the semi-discrete setting (Sadhu et al., 2024; Hundrieser et al., 2024b; del Barrio et al., 2024). For OT barycenters based on empirical measures significantly less is known, though recently some progress has been made in the context of finitely supported measures (Heinemann et al., 2022). Extending such results to general measures is not obvious and requires the need for alternative statistical modelling.

1.3 Contributions: Statistical Models and Deviation Bounds

The key contributions are summarized as follows. First, we propose three prototypical statistical models for general measures on finite spaces: the *multinomial model*, the *Bernoulli model*, and the *Poisson intensity model*. Each model gives rise to a specific measure estimator. For each model, we then analyze the statistical performance of plug-in estimators in approximating one of the following four quantities:

- (A) The population measure with respect to the (p, C)-Kantorovich-Rubinstein distance and in total variation norm (Section 2).
- (B) The (p, C)-Kantorovich-Rubinstein distance between two population measures in mean absolute deviation (Section 2).
- (C) The unbalanced optimal transport plan between two population measures with respect to the Hausdorff distance induced by the total variation norm (Section 3).
- (D) The (p, C)-barycenter of a finite collection of population measures in terms of the excess Fréchet function value and (p, C)-Kantorovich-Rubinstein distance (Section 4).

In addition to our theoretical analysis we show how our results can be employed for randomized computation with statistical guarantees (Section 5). Finally, we perform various simulations which showcase the performance of the introduced empirical estimators for different quantities above (Section 6).

1.3.1 Statistical models

In the following, we formalize the three statistical models and provide specific motivations for each. For an illustration of different realizations of these estimators see Figure 3.

Multinomial Model

For the *multinomial model*, we consider i.i.d. random variables $X_1, \dots, X_N \sim \frac{\mu}{\mathbb{M}(\mu)}$, where the total intensity $\mathbb{M}(\mu)$ is assumed to be known and strictly positive. The corresponding

unbiased empirical estimator is then defined as

$$\widehat{\mu}_N \coloneqq \frac{\mathbb{M}(\mu)}{N} \sum_{x \in \mathcal{X}} |\{k \in \{1, \dots, N\} \mid X_k = x\}| \, \delta_x. \tag{6}$$

This estimator is exactly the standard empirical measure associated to $\frac{\mu}{\mathbb{M}(\mu)}$ but rescaled with $\mathbb{M}(\mu)$. Insofar, this model extends the classical sampling approach for probability measures to measures of positive mass. A key motivation for introducing this model is resampling for randomized computation of UOT quantities. In real world data analysis it is common to encounter data (e.g., high-resolution images) which are out of reach for current state-of-the-art OT solvers. One idea in this scenario is to replace each measure by its empirical version and then compute the respective UOT quantity between these surrogates. For probability measures, this was introduced for the p-Wasserstein distance (Sommerfeld et al., 2019) and the p-Wasserstein barycenter (Heinemann et al., 2022). Statistical deviation bounds allow balancing computational complexity and approximation accuracy in terms of the sample size N. For more details on randomized computation in the present context we refer to Section 5.

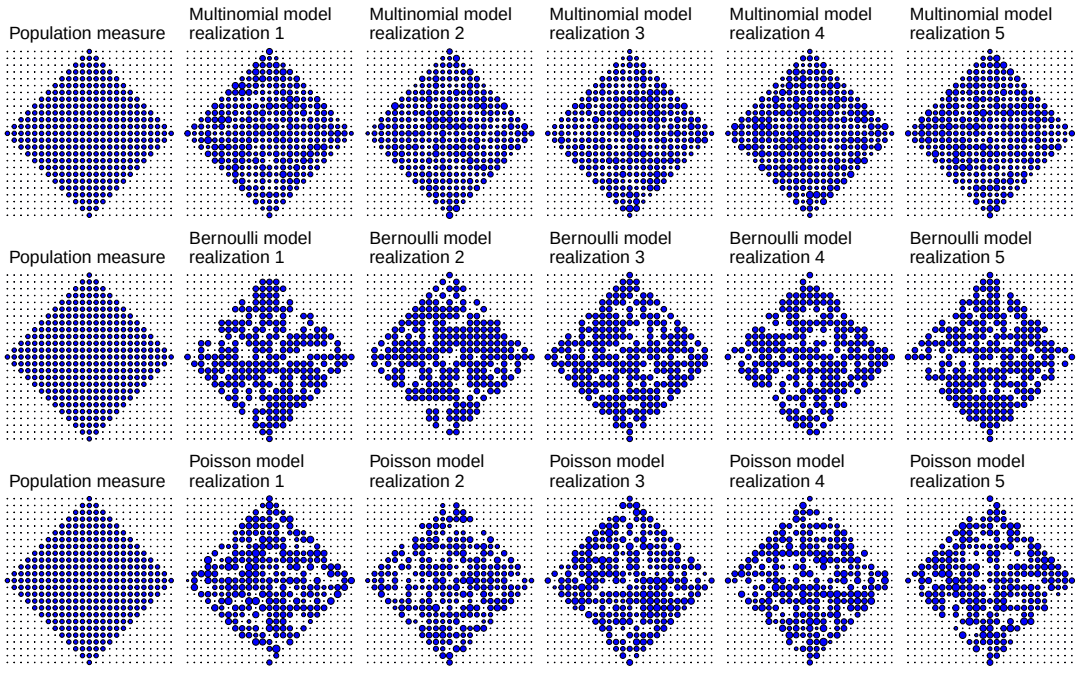


Figure 3: Realizations of measure estimators for the population measure (left column) based on the multinomial model with $N = 2000$ (top), Bernoulli model with homogeneous thinning $s = 0.75$ (middle) and Poisson intensity model with $s = 0.75$ and $t = 5$ (bottom). The thickness of each circle represents the mass assigned to the respective point.

Bernoulli Model

For the *Bernoulli model* we consider measures μ with $\mu(x) = 1$ for all $x \in \operatorname{supp}(\mu)$. Thus, the measure μ represents a point cloud in the ambient space $\mathcal{Y}$ with the total mass being the cardinality of the point cloud. We restrict the model to this setting, but we stress that generalizations to arbitrary masses at the individual locations are straightforward. For each location $x \in \mathcal{X}$ we assume to observe independent Bernoulli random variables $B_x \sim \mathrm{Ber}(s_x \mu(x))$ with a fixed *success probability* $s_x \in (0, 1]$. We denote $s_{\mathcal{X}} := (s_{x_1}, \dots, s_{x_M})$ where $x_1, \dots, x_M$ is an enumeration of all elements of $\mathcal{X}$, $|\mathcal{X}| = M$ and refer to $s_{\mathcal{X}}$ as *success vector*. A suitable unbiased estimator for μ is defined by

$$\widehat{\mu}_{s_{\mathcal{X}}} := \sum_{x \in \mathcal{X}} \frac{B_x}{s_x} \delta_x. \tag{7}$$

This estimator models a potentially *spatio-heterogenous* thinning of the measure μ in terms of Bernoulli random variables such that the total mass of $\widehat{\mu}_{s_{\mathcal{X}}}$ is close (but not always equal) to the total mass of μ. Notably, the Bernoulli field $(B_x)_{x \in \mathcal{X}}$ is a prototypical model for incomplete data, where data is missing at random and the Bernoulli variables serve as labels for this. It further arises, e.g., in the context of generalized linear models where the regressor X is linked to B_x by a link function in a non-parametric fashion. The Bernoulli model also underlies the sampling scheme for the subsequent Poisson intensity model, which occurs, e.g., in various imaging devices, such as fluorescence cell microscopy. There, fluorescent markers are, e.g., chemically attached to each protein within a complex protein ensemble and then are excited with a laser beam. The resulting, emitted photons indicate the spatial position of the objects of interest in the proper experimental setup (Kulaitis et al., 2021). However, the marker has a limited *labelling efficiency* $s_x \in (0, 1]$ at each location $x \in \mathcal{X}$, and we only observe a location which has been labelled by the marker and finally emits photons. We refer to Aspelmeier et al. (2015) for the further discussion on statistical aspects of high-resolution fluorescence microscopy.

Poisson Intensity Model

For the *Poisson intensity model* we fix a parameter $t > 0$ and a success probability $s \in (0, 1]$. Consider a collection of $|\mathcal{X}|$ independent Poisson random variables $P_x \sim \mathrm{Poi}(t\mu(x))$ with intensity $t\mu(x)$ at each location $x \in \mathcal{X}$ and independently of them Bernoulli random variables $B_x \sim \mathrm{Ber}(s)$ for each $x \in X$. A suitable unbiased estimator for μ is defined by

$$\widehat{\mu}_{t,s} := \frac{1}{st} \sum_{x \in \mathcal{X}} B_x P_x \delta_x. \tag{8}$$

In contrast to the Bernoulli model, in this model the success probability is assumed to be homogeneous (as opposed to the inhomogeneous probabilities in the Bernoulli model, though such generalizations are straightforward) and the values of the population measures at each support point are not necessarily equal to one. Hence, we have two independent layers of randomness in the construction of this empirical measure. First, we draw a location x with a certain probability s, then we observe random photon counts driven by a Poisson

distribution based on the mass of μ and the value of t.
This model is motivated by various tasks in photonic imaging, for example, fluorescence microscopy, X-ray imaging and positron emission tomography (PET), see Munk et al. (2020) for a survey. The finite space $\mathcal{X}$ represents the center of bins of a detection interface used to measure the emitted photons. The value $\mu(x)$ corresponds to the integrated underlying photon intensity over its respective bin. This intensity is proportional to an external source, such as a laser duration in fluorescence microscopy and modelled by the parameter $t > 0$. The Bernoulli random variable B_x models the possibility that in the bin of x a photon can not be recorded. This might be due to various effects that cause thinning, such as limited labelling efficiency, dead time of cameras or a loss of photons due to sparse detector tubes. The value of P_x corresponds to the number of photons which have been measured at the bin of x. Note that besides B_x there might be also additional effects present which do not disable the whole bin, but just prevent a single photon from being measured. All this causes a thinning of the process and is incorporated in the probability $s' \in (0, 1]$ that a single photon at any bin and any point in time can not be measured. In this case, the model can be shown to be equivalent to a Poisson model with parameter $ts' > 0$ instead of t (Aspelmeier et al., 2015), and is thus a special case of the general Poisson intensity model.

1.3.2 Summary of Statistical Deviation Bounds

Concerning approximation of population measure by its empirical counterpart, we show in Section 2 that there exist constants $\mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C), \mathcal{E}^{\mathrm{Ber}}_{p,\mathcal{X},\mu}(C), \mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C)$ such that for any measure μ and its estimator $\widehat{\mu}$, with $p \geq 1$, in each of the three statistical models it holds,

$$\mathbb{E}\left[\mathrm{KR}_{p,C}\left(\widehat{\mu}, \mu\right)\right] \leq \begin{cases} \mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C)^{\frac{1}{p}}\, N^{-\frac{1}{2p}}, & \text{if } \widehat{\mu} = \widehat{\mu}_N, \quad \text{(Multinomial)} \\ \mathcal{E}^{\mathrm{Ber}}_{p,\mathcal{X},\mu}(C)^{\frac{1}{p}}\, \psi(s_{\mathcal{X}})^{\frac{1}{p}}, & \text{if } \widehat{\mu} = \widehat{\mu}_{s_{\mathcal{X}}}, \quad \text{(Bernoulli)} \\ \mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C)^{\frac{1}{p}}\, \phi(t,s)^{\frac{1}{p}}, & \text{if } \widehat{\mu} = \widehat{\mu}_{t,s}, \quad \text{(Poisson)} \end{cases} \tag{9}$$

where for the *Multinomial model* we obtain a scaling rate of $N^{-\frac{1}{2p}}$, for the *Bernoulli model*

$$\psi(s_{\mathcal{X}}) = \begin{cases} \left(2\sum_{x\in\mathcal{X}}(1-s_x)\right), & \text{if } C \leq d_{\min} \coloneqq \min_{x\neq x'} d(x,x'), \\ \left(\sum_{x\in\mathcal{X}} \frac{1-s_x}{s_x}\right)^{\frac{1}{2}}, & \text{else,} \end{cases}$$

and for *Poisson intensity model*

$$\phi(t,s) = \begin{cases} \left(2(1-s)\mathbb{M}(\mu) + \frac{s}{\sqrt{t}}\sum_{x\in\mathcal{X}}\sqrt{\mu(x)}\right), & \text{if } C \leq d_{\min}, \\ \left(\frac{1}{st}\mathbb{M}(\mu) + \frac{1-s}{s}\sum_{x\in\mathcal{X}}\mu(x)^2\right)^{\frac{1}{2}}, & \text{else.} \end{cases}$$

Notably, in the multinomial model for $N \to \infty$, in the Poisson model for $t \to \infty$, $s \to 1$ and in the Bernoulli model for $s_{\mathcal{X}} \to \mathbf{1}_{\mathcal{X}}$, these upper bounds tend to zero. Our approach enables an explicit characterization of the constants $\mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C), \mathcal{E}^{\mathrm{Ber}}_{p,\mathcal{X},\mu}(C), \mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C)$ in terms of structural properties of the measures and space, such as total mass intensity and covering numbers (for details see Section 2 and Appendices A and B). We believe this to be particularly relevant for statistical tasks surrounding the KRD. As an example, we comment on the behavior of the constants if $\mathcal{X}$ is a subset of $\mathbb{R}^d$ equipped with Euclidean metric.

Example 1 (Compact ground space in $\mathbb{R}^D$) *For finitely supported measures on the unit ball in $\mathbb{R}^D$ with $|\mathcal{X}|$ support points and the Euclidean distance as the metric, it holds for $p = 2$ and $D \geq 1$ up a universal constant*[6] *(see Section 2.4 and Appendices A and B) that*

$$\frac{\mathcal{E}^{\mathrm{Mult}}_{2,\mathcal{X},\mu}(C)}{\mathbb{M}(\mu) + \mathbb{M}(\mu)^{1/2}}, \mathcal{E}^{\mathrm{Ber}}_{2,\mathcal{X},\mu}(C), \mathcal{E}^{\mathrm{Poi}}_{2,\mathcal{X},\mu}(C) \lesssim \begin{cases} C^2, & \text{if } C \leq d_{\min} \text{ or } D < 4, \\ C^2 + \log_2(|\mathcal{X}|), & \text{if } C > d_{\min} \text{ and } D = 4, \\ C^2 + |\mathcal{X}|^{\frac{1}{2} - \frac{2}{D}}, & \text{if } C > d_{\min} \text{ and } D > 4. \end{cases}$$

In particular, for $D < 4$ the constants are independent of the cardinality of the space $|\mathcal{X}|$, for $D = 4$ the dependence is logarithmic and for $D > 4$ there is a polynomial dependency in $|\mathcal{X}|$. We note however that the upper bound adapt to the intrinsic dimension of the domain on which the finite measure is supported, e.g., if a measure is supported on a $D' < D$ dimensional subspace of $[0,1]^D$ (e.g., a submanifold), then in the upper bound the dependency reduces to D' and where the suppressed constant depends on the domain.

In addition, we also establish related convergence statements with respect to the total variation norm (Theorem 4). Controlling the empirical estimator with respect to the (p, C)-KRD and total variation norm enables us to draw conclusions on the performance of plug-in estimators for the (p, C)-KRD. Indeed, by reverse triangle inequality and our stability bound in Section 2, Lemma 3 it holds

$$\begin{aligned} \mathbb{E}\left[|\mathrm{KR}_{p,C}(\widehat{\mu}, \widehat{\nu}) - \mathrm{KR}_{p,C}(\mu, \nu)|\right] \leq \min \Big(& \mathbb{E}\left[\mathrm{KR}_{p,C}(\widehat{\mu}, \mu)\right] + \mathbb{E}\left[\mathrm{KR}_{p,C}(\widehat{\nu}, \nu)\right], \\ & 2C^p \mathrm{KR}^{1-p}_{p,C}(\mu, \nu) \left(\mathbb{E}\left[\mathrm{TV}(\widehat{\mu}, \mu)\right] + \mathbb{E}\left[\mathrm{TV}(\widehat{\nu}, \nu)\right]\right) \Big), \end{aligned} \tag{10}$$

where $\mathrm{TV}(\mu, \nu) = \sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)|$ is the *total variation distance*. This asserts that the $\mathrm{KR}_{p,C}(\mu, \nu)$ is well-approximated by $\mathrm{KR}_{p,C}(\widehat{\mu}, \widehat{\nu})$ as soon as the empirical estimators $\widehat{\mu}$ and $\widehat{\nu}$ approximate the population measures well (Corollary 7 and Theorem 8).

In addition to the UOT *values* as in (10), we also establish a novel quantitative stability bound for UOT *plans* with parameters $p \geq 1$ and $C > 0$. Since UOT plans are not necessarily unique, our stability result is stated for the collection of UOT plans $\mathbf{P}^*_{p,C}(\mu, \nu)$ and is quantified in terms of the Hausdorff distance $\mathcal{H}_{\mathrm{TV}}$ induced by the total variation norm (Theorem 11),

$$\begin{aligned} & \mathcal{H}_{\mathrm{TV}}\left(\mathbf{P}^*_{p,C}(\widehat{\mu}, \widehat{\nu}), \mathbf{P}^*_{p,C}(\mu, \nu)\right) \\ & \leq 4\left(|\mathcal{X}| + 1\right)\left(\mathrm{TV}(\widehat{\mu}, \mu) + |\mathbb{M}(\widehat{\mu}) - \mathbb{M}(\mu)| + \mathrm{TV}(\widehat{\nu}, \nu) + |\mathbb{M}(\widehat{\nu}) - \mathbb{M}(\nu)|\right). \end{aligned}$$

Based on this bound, we can make quantitative statements about the performance of plug-in estimators for UOT plans (Theorem 12). Notably, we also establish a quantitative stability bound for balanced OT.

Finally, the convergence analysis of empirical estimators with respect to the (p, C)-KRD also enables us to draw conclusions for (p, C)-barycenters (defined in (3)), see Theorems 16 and 17. To elaborate, let μ^* be any (p, C)-barycenter of the population measures $\mu^1, \ldots, \mu^J$

6. We write $A \lesssim B$ if there exists a universal constant $\kappa > 0$ such that $A \leq \kappa B$.

and let $\widehat{\mu}^\star$ be any (p, C)-barycenter of the empirical measures $\widehat{\mu}^1, \ldots, \widehat{\mu}^J$. Since neither $\mu^\star$ nor $\widehat{\mu}^\star$ is necessarily unique, we quantify the error of the empirical counterpart $\widehat{\mathbf{B}}^\star$ in approximating the population set $\mathbf{B}^\star$ via (recall (3))

$$F_{p,C}(\widehat{\mu}^\star) - F_{p,C}(\mu^\star) \quad \text{and} \quad \sup_{\widehat{\mu}^\star \in \widehat{\mathbf{B}}^\star} \inf_{\mu^\star \in \mathbf{B}^\star} \mathrm{KR}_{p,C}(\mu^\star, \widehat{\mu}^\star). \tag{11}$$

Notably, the term on the right-hand side quantifies the maximal difference between an empirical barycenter to the closest population barycenter. Although this is a slightly weaker result than the analysis in terms of the Hausdorff distance, it is sufficient for practical considerations, as it details how well the "worst choice" for the empirical barycenters approximates its population counterpart. Both expressions in (11) can be related to the (p, C)-KRD error between empirical and population measure in (9), which enables us to quantify the performance of empirical barycenters.

2. Empirical Kantorovich-Rubinstein Distances

In this section we investigate the Poisson model and analyze how fast the empirical measure estimator tends to its population counterpart in terms of the (p, C)-Kantorovich-Rubinstein distance and the total variation norm. These bounds allow us to quantify how fast the plug-in estimator tends to the population (p, C)-Kantorovich-Rubinstein distance. Results for the multinomial and Bernoulli model follow along the same reasoning. Corresponding deviation bounds and proofs are provided in Appendices A and B.

2.1 Stability Bounds for Kantorovich-Rubinstein Distance

The convergence results of this section are based on novel stability bounds for the UOT distance, which we detail in the following. The first is based on a tree approximation of the space $\mathcal{X}$ (Lemma 1), whereas the second relies on the dual formulation of the UOT cost (Lemma 3). The proofs for this section are detailed in Appendix D.1.

2.1.1 Stability Bound via Tree Approximation of Domain

Let $\mathcal{T} = (V, E)$ be a rooted, ultrametric tree with height function $h : V \longrightarrow \mathbb{R}_+$ and root r. For two nodes $\mathsf{u}, \mathsf{v} \in V$, denote the unique path between u and v in $\mathcal{T}$ by $\mathcal{P}(\mathsf{u}, \mathsf{v})$. For a node $\mathsf{v} \in V$ its *children* are the elements of the set $\mathcal{C}(\mathsf{v}) = \{\mathsf{w} \in V \mid \mathsf{v} \in \mathcal{P}(\mathsf{w}, \mathsf{r})\}$. The *parent* $par(\mathsf{v})$ of a node v is the unique node with $(par(\mathsf{v}), \mathsf{v}) \in E$ and $h(\mathsf{v}) < h(par(\mathsf{v}))$. For any $C > 0$, define the set

$$\mathcal{R}(C) \coloneqq \{\mathsf{v} \in V \mid h(\mathsf{v}) \leq C/2 < h(\mathrm{par}(\mathsf{v}))\} \tag{12}$$

with the convention that $\mathcal{R}(C) = \{\mathsf{r}\}$ if $\frac{C}{2} \geq h(\mathsf{r})$. The goal is to control the (p, C)-KRD on the finite metric space $(\mathcal{X}, d)$ by bounding it from above by a dominating distance $d_{\mathcal{T}}$ induced[7] from a tree $\mathcal{T}$ with the elements of $\mathcal{X}$ as vertices and a height function h such

7. For two vertices of the tree $\mathcal{T}$, we define their distance $d_{\mathcal{T}}$ as the sum of the weights of the edges included in the unique path between the two vertices. Here, the weight of an edge joining two vertices v and $par(v)$ is given by $h(par(v)) - h(v)$.

that $d(x,x') \leq d_{\mathcal{T}}(x,x')$. In this case and by the definition of the Kantorovich-Rubinstein distance it holds for all measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ that

$$\mathrm{KR}_{p,C}(\mu,\nu) \leq \mathrm{KR}_{d_{\mathcal{T}}^p,C}(\mu,\nu), \tag{13}$$

where $\mathrm{KR}_{d_{\mathcal{T}}^p,C}(\mu,\nu)$ denotes the (p,C)-KRD w.r.t. the ground space $(\mathcal{X}, d_{\mathcal{T}})$. Moreover, if $\mathcal{T}$ is an ultrametric tree with leaf nodes L and height function $h\colon V \to \mathbb{R}_+$ inducing[7] the tree metric $d_{\mathcal{T}}$ and the two measures $\mu^L, \nu^L \in \mathcal{M}_+(L)$ supported on the leaf nodes of $\mathcal{T}$, then it holds (Heinemann et al., 2023, Theorem 2.3) that

$$\begin{aligned}
&\mathrm{KR}^p_{d_{\mathcal{T}}^p,C}\left(\mu^L, \nu^L\right) = \\
&\sum_{\mathsf{v} \in \mathcal{R}(C)} \Bigg(2^{p-1} \sum_{\mathsf{w} \in \mathcal{C}(\mathsf{v}) \setminus \{\mathsf{v}\}} \Big(\left(h(par(\mathsf{w}))^p - h(\mathsf{w})^p\right) \left|\mu^L(\mathcal{C}(\mathsf{w})) - \nu^L(\mathcal{C}(\mathsf{w}))\right| \Big) \\
&+ \left(\frac{C^p}{2} - 2^{p-1} h(\mathsf{v})^p \right) \left|\mu^L(\mathcal{C}(\mathsf{v})) - \nu^L(\mathcal{C}(\mathsf{v}))\right| \Bigg).
\end{aligned} \tag{14}$$

The construction of $\mathcal{T}$, such that (13) holds, is as follows. Fix some *depth level* $L \in \mathbb{N}$. For some *resolution* $q > 1$ and level $j = 0, \ldots, L$ define the covering set[8] $Q_j \coloneqq \mathcal{N}(\mathcal{X}, q^{-j}\mathrm{diam}(\mathcal{X})) \subset \mathcal{X}$ and let $Q_{L+1} \coloneqq \mathcal{X}$. Any point $x \in Q_j$ is considered as a node at level j of a tree $\mathcal{T}$ and denoted as (x,j) to emphasize its level position. An illustration of this approximation is given in Figure 4. For level $j = 0$ this yields a single element in Q_0 which serves as the root of the tree. For $j = 0, \ldots, L$ a node (x,j) at level j is connected to one node $(x', j+1)$ at level $j+1$ if their distance satisfies $d(x,x') \leq q^{-j}\mathrm{diam}(\mathcal{X})$ (ties are broken arbitrarily). The edge weight of the corresponding edge is set equal to $q^{-j}\mathrm{diam}(\mathcal{X})$. Consequently, the height of each node only depends on its assigned level $0 \leq l \leq L+1$ and is defined as $h_{q,L}\colon \{0, \ldots, L+1\} \to \mathbb{R}$ by [9]

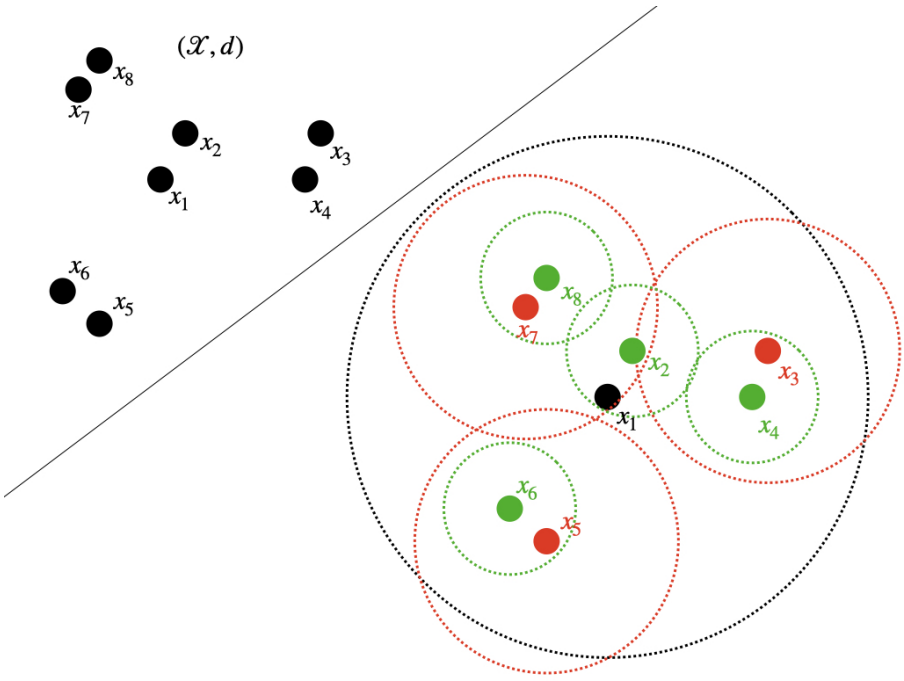
$$h_{q,L}(l) = \sum_{j=l}^{L} q^{-j}\mathrm{diam}(\mathcal{X}) = \frac{q^{1-l} - q^{-L}}{q-1}\mathrm{diam}(\mathcal{X}). \tag{15}$$

By definition the space $\mathcal{X}$ is embedded in level $L+1$ as the leaf nodes of $\mathcal{T}$ with height $h_{q,L}(L+1) = 0$. By a straightforward computation it holds for two points $x, x' \in \mathcal{X}$ considered as embedded in $\mathcal{T}$ as $(x, L+1)$ and $(x', L+1)$ that

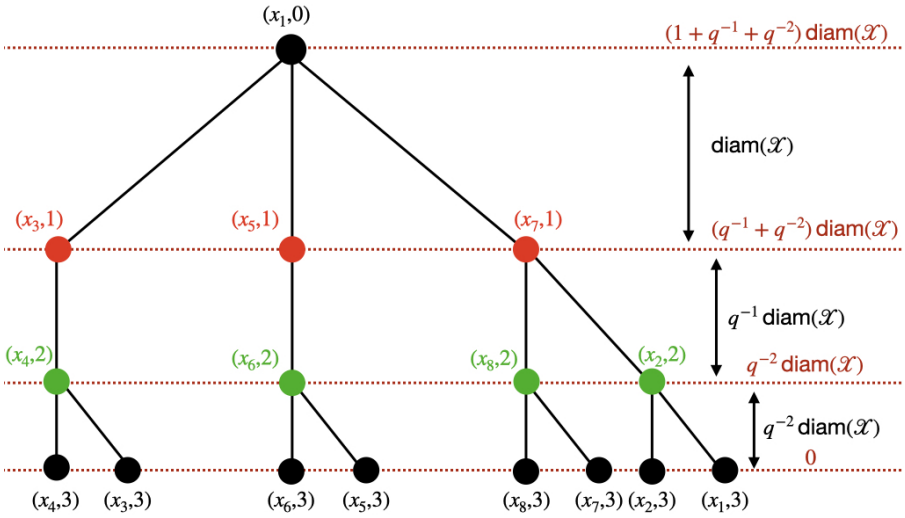
$$d^p(x,x') \leq d^p_{\mathcal{T}}((x, L+1), (x', L+1)). \tag{16}$$

8. For a metric space $(\mathcal{X}, d)$ an ε-cover is a set of points $\{x_1, \ldots, x_m\} \subset \mathcal{X}$ such that for each $x \in \mathcal{X}$, there exists some $1 \leq i \leq m$ such that $d(x, x_i) \leq \varepsilon$. The smallest such set is denoted as $\mathcal{N}(\mathcal{X}, \varepsilon)$.

9. The construction of the ultra-metric tree is based on approximating the underlying domain at varying precisions $(\delta_0, \ldots, \delta_L) = (q^{-l}\,\mathrm{diam}(\mathcal{X}))_{l=0,\ldots,L}$ with as few points as possible. If one were to select the precisions differently, the height function would change to $h(l) \coloneqq \sum_{j=l}^{L} \delta_j$, possibly leading to different convergence statements. Our choice is inspired by that of Dereich et al. (2013) and Fournier and Guillin (2015), who derived sharp rates of convergence for the empirical Wasserstein distance in Euclidean settings. Based on this choice, we establish in Section 2.4 convergence results for different regimes which align with those of previous articles.



(a)



(b)

Figure 4: **Ground metric approximation by an ultrametric tree distance: (a)** A finite metric space $(\mathcal{X}, d)$ and its covering sets Q_0 (black), Q_1 (red) and Q_2 (green) for $L = 2$. **(b)** Based on the covering sets from (a) an ultrametric tree is constructed. The metric space $\mathcal{X}$ is embedded in level $L + 1 = 3$ and equal to all leaf nodes of that tree.

The measures μ, ν are embedded into $\mathcal{T}$ as measures μ^L, ν^L supported only on leaf nodes of $\mathcal{T}$, and thus it follows from (16) that

$$\mathrm{KR}_{p,C}(\mu,\nu) \leq \mathrm{KR}_{d_{\mathcal{T}}^p,C}\left(\mu^L,\nu^L\right).$$

In combination with the closed formula from (14) this yields an upper bound on the (p,C)-KRD. Whenever clear from the context the notation is alleviated by writing $\mathsf{v} \in Q_l$ instead of $(\mathsf{v},l) \in Q_l$.

Lemma 1 *Let $(\mathcal{X},d)$ be a finite metric space and let $\mu,\nu \in \mathcal{M}_+(\mathcal{X})$ with total mass $\mathbb{M}(\mu)$ and $\mathbb{M}(\nu)$, respectively. Let $p \geq 1$ and $C > 0$. Then for any resolution $q > 1$, depth $L \in \mathbb{N}$ and height function (15) with*

$$h_{q,L}(k) = \frac{q^{1-k} - q^{-L}}{q-1}\mathrm{diam}(\mathcal{X})$$

it holds that

$$\mathrm{KR}_{p,C}^p(\mu,\nu) \leq \begin{cases} \left(\frac{C^p}{2} - 2^{p-1}h_{q,L}(0)^p\right)|\mathbb{M}(\mu)-\mathbb{M}(\nu)| \\ +B_{q,p,L,\mathcal{X}}(1), \\ \qquad\qquad \text{if } C \geq 2h_{q,L}(0), \\ B_{q,p,L,\mathcal{X}}(l-1), \\ \qquad\qquad \text{if } 2h_{q,L}(l) \leq C < 2h_{q,L}(l-1), \\ \frac{C^p}{2}\mathrm{TV}(\mu,\nu), \\ \qquad\qquad \text{if } C \leq \left(2h_{q,L}(L) \vee \min_{x\neq x'} d(x,x')\right), \end{cases}$$

where we define

$$B_{q,p,L,\mathcal{X}}(l) \coloneqq 2^{p-1}\sum_{j=l}^{L+1}\sum_{x\in Q_j}\left(h_{q,L}(j-1)^p - h_{q,L}(j)^p\right)\left|\mu^L(\mathcal{C}(x)) - \nu^L(\mathcal{C}(x))\right|.$$

Remark 2 *According to Theorem 1, for any two measures $\mu,\nu \in \mathcal{M}_+(\mathcal{X})$ and $C \leq \left(2h_{q,L}(L) \vee \min_{x\neq x'} d(x,x')\right)$ it always holds that $\mathrm{KR}_{p,C}^p(\mu,\nu) \leq \frac{C^p}{2}\mathrm{TV}(\mu,\nu) \leq \frac{C^p}{2}$. In particular, for $C \searrow 0$ this implies that*

$$\sup_{\mu,\nu\in\mathcal{M}_+(\mathcal{X})}\frac{\mathrm{KR}_{p,C}^p(\mu,\nu)}{1+\mathbb{M}(\mu+\nu)} \leq \frac{C^p}{2}\frac{\mathrm{TV}(\mu,\nu)}{1+\mathbb{M}(\mu+\nu)} \leq \frac{C^p}{2} \searrow 0.$$

2.1.2 Stability Bound via Dual Formulation

The stability bound based on the tree construction from previous subsection yields sharp statements about the convergence of the measure estimator to its population counterpart.

By triangle inequality, this yields sharp convergence rates for plug-in estimators of the KRD when population measures are close, but they are suboptimal when the population measures are strictly separated. In this subsection we follow a different approach to derive a stability bound for the KRD which is inspired by Chizat et al. (2020) and Manole and Niles-Weed (2024) and based on the dual representation of the KRD.

Lemma 3 *Let $(\mathcal{X}, d)$ be a finite metric space and consider two pairs of measures $\mu^1, \nu^1, \mu^2, \nu^2 \in \mathcal{M}_+(\mathcal{X})$. Let $p \geq 1$ and $C > 0$. Then, it follows that*

$$|\mathrm{KR}^p_{p,C}(\mu^1, \nu^1) - \mathrm{KR}^p_{p,C}(\mu^2, \nu^2)| \leq \frac{C^p}{2} \begin{cases} 4\left(\mathrm{TV}(\mu^1, \mu^2) + \mathrm{TV}(\nu^1, \nu^2)\right), \\ \qquad \text{if } C > \min\{\Delta(\mu^1, \nu^1), \Delta(\mu^2, \nu^2)\}, \\ \left(\left|\mathbb{M}(\mu^1) - \mathbb{M}(\mu^2)\right| + \left|\mathbb{M}(\nu^1) - \mathbb{M}(\nu^2)\right|\right), \\ \qquad \text{else,} \end{cases} \tag{17}$$

where we define

$$\Delta(\mu, \nu) := \begin{cases} \min_{x \in \mathrm{supp}(\mu), x' \in \mathrm{supp}(\nu)} d(x, x'), & \text{if } \mu \neq 0 \text{ and } \nu \neq 0, \\ +\infty, & \text{if } \mu = 0 \text{ or } \nu = 0. \end{cases}$$

Moreover, under $\mathrm{KR}_{p,C}(\mu^1, \nu^1) \geq \delta > 0$ it holds that

$$|\mathrm{KR}_{p,C}(\mu^1, \nu^1) - \mathrm{KR}_{p,C}(\mu^2, \nu^2)| \leq \delta^{1-p} |\mathrm{KR}^p_{p,C}(\mu^1, \nu^1) - \mathrm{KR}^p_{p,C}(\mu^2, \nu^2)|.$$

2.2 ($\mathbf{p}, \mathbf{C}$)-Kantorovich-Rubinstein Deviation Bound

In this section we first derive statistical deviation bounds for approximation error empirical estimators for their population counterpart with respect to the TV norm and (p, C)-KRD. We then continue by analyzing the performance of plug-in estimators. The omitted proofs of the following Theorems 4 and 5 can be found in Appendix D.2.

Theorem 4 (Expected deviation of estimator in TV and KRD) *Let $(\mathcal{X}, d)$ be a finite metric space and $\mu \in \mathcal{M}_+(\mathcal{X})$ with total mass $\mathbb{M}(\mu)$. Let $\widehat{\mu}_{t,s}$ be the estimator from (8). Then, for any $p \geq 1$ and $C > 0$ it holds that*

$$\mathbb{E}\left[\mathrm{KR}^p_{p,C}\left(\widehat{\mu}_{t,s}, \mu\right)\right] \leq C^p \mathbb{E}\left[\mathrm{TV}\left(\widehat{\mu}_{t,s}, \mu\right)\right] \leq C^p \left(2(1-s)\mathbb{M}(\mu) + \frac{s}{\sqrt{t}} \sum_{x \in \mathcal{X}} \sqrt{\mu(x)}\right).$$

Theorem 5 (Expected deviation of estimator in KRD) *Let $(\mathcal{X}, d)$ be a finite metric space and $\mu \in \mathcal{M}_+(\mathcal{X})$. Let $\widehat{\mu}_{t,s}$ be the estimator from (8). Then, for any $p \geq 1$, $C > 0$,*

resolution $q > 1$ and depth $L \in \mathbb{N}$ it holds that

$$\mathbb{E}\left[\mathrm{KR}_{p,C}\left(\widehat{\mu}_{t,s}, \mu\right)\right] \leq \mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C,q,L)^{1/p} \begin{cases} \left(2(1-s)\mathbb{M}(\mu) + \frac{s}{\sqrt{t}} \sum_{x\in\mathcal{X}} \sqrt{\mu(x)}\right)^{\frac{1}{p}}, \\ \qquad\qquad \text{if } C \leq \left(2h_{q,L}(L) \vee \min_{x\neq x'} d(x,x')\right), \\ \left(\frac{1}{st}\mathbb{M}(\mu) + \frac{1-s}{s}\sum_{x\in\mathcal{X}} \mu(x)^2\right)^{\frac{1}{2p}}, \\ \qquad\qquad \text{else.} \end{cases}$$

For

$$A_{q,p,L,\mathcal{X}}(l) := \mathrm{diam}(\mathcal{X})^p 2^{p-1} \left(q^{-Lp} |\mathcal{X}|^{\frac{1}{2}} + \left(\frac{q}{q-1}\right)^p \sum_{j=l}^{L} q^{p-jp} |Q_j|^{\frac{1}{2}} \right),$$

the constant is equal to

$$\mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C,q,L) = \begin{cases} \left(\frac{C^p}{2} - 2^{p-1}\left(\frac{q-q^{-L}}{q-1}\mathrm{diam}(\mathcal{X})\right)^p\right) + A_{q,p,L,\mathcal{X}}(1), \\ \qquad\qquad \text{if } C \geq 2h_{q,L}(0), \\ A_{q,p,L,\mathcal{X}}(l), \\ \qquad\qquad \text{if } 2h_{q,L}(l) \leq C < 2h_{q,L}(l-1), \\ \frac{C^p}{2}, \\ \qquad\qquad \text{if } C \leq \left(2h_{q,L}(L) \vee \min_{x\neq x'} d(x,x')\right), \end{cases}$$

Furthermore, for $p = 1$ the factor $\frac{q}{(q-1)}$ in $A_{q,1,L,\mathcal{X}}(l)$ can be removed for all $l = 1, \ldots, L$.

The constant $\mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C,q,L)$ is reminiscent of the constants for similar deviation bounds for optimal transport between finitely supported measures (Boissard and Le Gouic, 2014; Sommerfeld et al., 2019). However, in the case of UOT one finds an interesting case distinction into roughly three cases depending on the relation between the penalty parameter C of the (p,C)-KRD and the resolution q and depth L of the tree approximation. The different constants arise from the fact that C controls the maximal range at which transport occurs in an UOT plan. In particular, if $d(x,x') > C^p$, then for any UOT plan π it holds $\pi(x,x') = 0$. If C is sufficiently large ($C \geq 2h(0)$), i.e., larger than the diameter of $\mathcal{X}$, then $\mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C,q,L)$ coincides with the deviation bounds for usual optimal transport, however, there is an additional summand arising from the necessary estimation of the true total mass $\mathbb{M}(\mu)$ (see also Lemma 1). For sufficiently small C, e.g., $C < \min_{x\neq x'} d(x,x')$, the (p,C)-KRD is proportional to the TV distance, hence we obtain a constant which is oblivious to the geometry of the ground space (see Theorem 4). For an intermediate value of C, the UOT problem on the ultra-metric tree $\mathcal{T}$ decomposes into smaller problems on subtrees of $\mathcal{T}$ (for details see the proof of (14) in Heinemann et al. (2023)) depending on C. The expected (p,C)-KRD error then depends on the size of these subtrees and the mass estimation error inherent to the total mass on these subtrees.

Remark 6 *Since the deviation bound holds for any resolution $q > 1$ and depth $L \in \mathbb{N}$ one can optimize and equivalently state upper bounds in terms of the infimum over those parameters. When the dependence on q or L is omitted, it is assumed that the infimum over those parameters has been taken, i.e.*

$$\mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C) = \inf_{L\in\mathbb{N},q>1} \mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C,q,L).$$

From the reverse triangle inequality we immediately obtain the following corollary from Theorem 5.

Corollary 7 (Expected deviation of empirical KRD) *Let $(\mathcal{X}, d)$ be a finite metric space and $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$. Let $\widehat{\mu}_{t,s}, \widehat{\nu}_{t,s}$ be the estimator from (8) for each of these measures, respectively. Then, for any $p \geq 1$, resolution $q > 1$ and depth $L \in \mathbb{N}$ it holds that*

$$\begin{aligned}
&\mathbb{E}\left[\left|\mathrm{KR}_{p,C}\left(\widehat{\mu}_{t,s}, \widehat{\nu}_{t,s}\right) - \mathrm{KR}_{p,C}\left(\mu,\nu\right)\right|\right] \\
&\leq \mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C,q,L)^{1/p} \begin{cases}
\left(2(1-s)\mathbb{M}(\mu+\nu) + \frac{s}{\sqrt{t}}\sum_{x\in\mathcal{X}}\left(\sqrt{\mu(x)}+\sqrt{\nu(x)}\right)\right)^{\frac{1}{p}}, \\
\qquad \textit{if } C < \min_{x\neq x'\in\mathcal{X}} d(x,x'), \\
\left(\frac{1}{st}\mathbb{M}(\mu+\nu) + \frac{1-s}{s}\sum_{x\in\mathcal{X}}(\mu(x)^2+\nu(x)^2)\right)^{\frac{1}{p}}, \\
\qquad \textit{else.}
\end{cases}
\end{aligned}$$

The deviation bound established in Theorem 7 is sharp in the sense that they are optimal up to constants for the regime where $\mu = \nu$ (see Section 2.3). Moreover, even if μ and ν differ but are permitted to be arbitrarily close, the convergence rates are sharp. However, if the measures are strictly separated, the rates are suboptimal, an observation which was previously made in the context of empirical optimal transport Chizat et al. (2020); Manole and Niles-Weed (2024). The subsequent deviation bound provides a refinement for the empirical KRD in this regime and is based on the stability bound via duality (Lemma 3).

Theorem 8 (Expected deviation of empirical KRD under separation) *Consider the same setting as in Corollary 7 and assume additionally that $\mathrm{KR}_{p,C}(\mu,\nu) \geq C\delta > 0$. Then it follows that*

$$\begin{aligned}
&\mathbb{E}\left[\left|\mathrm{KR}_{p,C}\left(\widehat{\mu}_{t,s}, \widehat{\nu}_{t,s}\right) - \mathrm{KR}_{p,C}\left(\mu,\nu\right)\right|\right] \\
&\leq \frac{\delta^{1-p}C}{2} \begin{cases}
4\left(2(1-s)\mathbb{M}(\mu+\nu) + \frac{s}{\sqrt{t}}\sum_{x\in\mathcal{X}}\left(\sqrt{\mu(x)}+\sqrt{\nu(x)}\right)\right)^{\frac{1}{2}}, \\
\qquad \textit{if } C > \min_{\substack{x\in\mathrm{supp}(\mu)\\ x'\in\mathrm{supp}(\nu)}} d(x,x'), \\
\left(\frac{1}{st}\mathbb{M}(\mu+\nu) + \frac{1-s}{s}\sum_{x\in\mathcal{X}}(\mu(x)^2+\nu(x)^2)\right)^{\frac{1}{2}}, \\
\qquad \textit{else.}
\end{cases}
\end{aligned}$$

Proof The result follows from Lemma 3 by plugging in the bounds of $\mathrm{TV}(\widehat{\mu}_{t,s}, \mu)$ and $\mathbb{E}[|\mathbb{M}(\widehat{\mu}_{t,s}) - \mathbb{M}(\mu)|]$, obtained in Theorem 4 and in (31). ■

In the formulation of the statement, we impose the condition that $\mathrm{KR}_{p,C}(\mu,\nu) \geq C\delta > 0$. This condition is motivated from the fact that $\mathrm{KR}_{p,C}(\mu,\nu) = 2^{-1/p} C \mathrm{TV}(\mu,\nu)^{1/p}$ for $C < \min_{x \neq x'} d(x,x')$. In particular, this way, we obtain a linear scaling in C in the upper bound of Theorem 8, which captures the correct scaling (see Section 2.3).

Remark 9 *To compare the established convergence results in Corollary 7 and Theorem 8, set $s = 1$ and note that if μ is close to ν, the convergence rate of the empirical KRD is of order $t^{-1/2p}$, whereas under a strict separation constraint the rate is instead of order $t^{-1/2}$. This is a consequence of the fact that the p-th root functional $x \mapsto x^{1/p}$ is Lipschitz-continuous at the origin if and only if $p = 1$.*

2.3 Rate Optimality

In the following, we provide some intuition and consequences of the deviation bound provided in Theorem 5. The term $\phi(s,t)$ must necessarily contain a sum of terms depending on s and t, respectively. In particular, neither choosing $s = 1$ nor letting t go to infinity for $s < 1$, would yield a zero error. For any fixed $t > 0$ and $s = 1$ the expected (p,C)-KRD error is clearly non-zero as the mass of the measures at each location is in general not estimated correctly. Similarly, for any fixed $s < 1$ letting $t \to \infty$ can not yield an expected (p,C)-KRD error of zero as on average $(1-s)|\mathcal{X}| > 0$ support points of μ are not observed. However, for $s = 1$ the error vanishes for $t \to \infty$, as we observe all support points of μ and then the strong law of large numbers guarantees the convergence of the weights at each location. It remains to verify whether the rate in t is optimal. For this, fix $s = 1$ and observe that

$$\min\{C, \min_{x \neq x'} d(x,x')\}^p \mathrm{TV}(\mu,\nu) \leq \mathrm{KR}^p_{p,C}(\mu,\nu) \leq C^p \mathrm{TV}(\mu,\nu). \tag{18}$$

To show that the rate $t^{-\frac{1}{2}}$ in Theorem 5 is sharp, we prove that $t^{\frac{1}{2}} \mathrm{TV}(\mu, \widehat{\mu}_{t,1})$ converges in distribution for $t \to \infty$ to a non-degenerate distribution. By combining the central limit theorem for Poisson random variables in conjunction with the continuous mapping theorem, it follows for $t \to \infty$ that

$$t^{\frac{1}{2}} \mathrm{TV}(\mu, \widehat{\mu}_{t,1}) = t^{\frac{1}{2}} \sum_{x \in \mathcal{X}} |\mu(x) - \widehat{\mu}_{t,1}(x)| = t^{-\frac{1}{2}} \sum_{x \in \mathcal{X}} |t\mu(x) - P_{x,t}| \xrightarrow{\mathcal{D}} \sum_{x \in \mathcal{X}} \sqrt{\mu(x)} |Z_x|,$$

where $P_{x,t} \sim Poi(\mu(x)t)$ and $Z_x \sim \mathcal{N}(0,1)$ with $x \in \mathcal{X}$ are jointly independent random variables. Hence, by the Portemanteau lemma for the function $x \mapsto x^{1/p}$, we conclude that

$$\liminf_{t \to \infty} \mathbb{E}[t^{\frac{1}{2p}} \mathrm{TV}(\mu, \widehat{\mu}_{t,1})^{\frac{1}{p}}] \geq \mathbb{E}\Big[\Big(\sum_{x \in \mathcal{X}} \sqrt{\mu(x)} |Z_x| \Big)^{\frac{1}{p}}\Big] > 0,$$

In conjunction with (18), the expectation $\mathbb{E}\left[\mathrm{KR}_{p,C}(\mu, \widehat{\mu}_{t,1})\right]$ is decreasing at least with order $Ct^{-\frac{1}{2p}}$ and the rate in t of Theorem 5 is thus sharp.

To illustrate sharpness of the convergence statement Theorem 8 in the regime of different measures, take $\nu = 0$, which yields $\widehat{\nu}_t = \nu = 0$ for all $t > 0$, and note that for every measure μ on $\mathcal{X}$ it follows that

$$\mathrm{KR}_{p,C}^p(\mu, \nu) = \frac{C^p}{2}\mathbb{M}(\mu).$$

Then, by the central limit theorem for Poisson random variables in conjunction with the delta method for $x \mapsto x^{1/p}$ and the continuous mapping theorem for the absolute value function, it follows if $\mathbb{M}(\mu) > 0$ for $t \to \infty$ that

$$t^{\frac{1}{2}} \left|\mathrm{KR}_{p,C}(\widehat{\mu}_{t,1}, \widehat{\nu}_{t,1}) - \mathrm{KR}_{p,C}(\mu, \nu)\right| = \frac{t^{\frac{1}{2}} C}{2^{1/p}} \left|\mathbb{M}(\mu)^{1/p} - \mathbb{M}(\widehat{\mu}_{t,1})^{1/p}\right| \xrightarrow{\mathcal{D}} \frac{C\mathbb{M}(\mu)^{-\frac{1}{2}+\frac{1}{p}}}{2^{1/p}}|Z|$$

for $Z \sim \mathcal{N}(0, 1)$. Again invoking the Portemanteau theorem thus yields that

$$\liminf_{t\to\infty} \mathbb{E}\left[t^{\frac{1}{2}} \left|\mathrm{KR}_{p,C}(\widehat{\mu}_{t,1}, \widehat{\nu}_{t,1}) - \mathrm{KR}_{p,C}(\mu, \nu)\right|\right] \geq \frac{C}{\sqrt{\pi}} \left(\frac{\mathbb{M}(\mu)}{2}\right)^{-\frac{1}{2}+\frac{1}{p}},$$

which shows that the convergence rate of order $Ct^{-1/2}$ in Theorem 8 is sharp.

2.4 Explicit Bounds for Euclidean Spaces

While the constants in the previous theorem are valid on arbitrary metric spaces, more explicit bounds can be derived for many practical applications. Thus, we assume that for the ε-covering number of $\mathcal{X}$, there exists a constant $A > 0$ such that

$$\mathcal{N}(\mathcal{X}, \varepsilon \cdot \mathrm{diam}(\mathcal{X})) \leq \min(A\varepsilon^{-\alpha}, |\mathcal{X}|) \quad \text{for all } \varepsilon \in (0, 1]. \tag{19}$$

This assumption covers, for instance, the setting when $\mathcal{X}$ is a finite subset of an α-dimensional submanifold of $\mathbb{R}^D$ and with d chosen as the Euclidean distance d_2. Notably, if the domain is the unit ball $\{x \in \mathbb{R}^D | \, \|x\| \leq 1\}$, then $\alpha = D$ and $A = 2$ are viable choices (Vershynin, 2018, Corollary 4.2.11). We fix $q = 2$ for simplicity. By repeating the argument within this framework, we can compute explicit upper bounds on the constants in Theorem 5. For $\alpha < 2p$ and $L \to \infty$, it holds

$$\mathcal{E}_{p,\mathcal{X},\mu}^{\mathrm{Poi}}(C) \leq \begin{cases} \frac{C^p}{2} - 2^{2p-1}\mathrm{diam}(\mathcal{X})^p + A^{1/2}\mathrm{diam}(\mathcal{X})^p 2^{3p-1}\frac{2^{\alpha/2-p}}{1-2^{\alpha/2-p}}, \\ \qquad\qquad \text{if } C \geq 2h_L(0), \\ A^{1/2}2^{3p-1}\mathrm{diam}(\mathcal{X})^p \frac{2^{(\alpha/2-p)l}}{1-2^{\alpha/2-p}}, \\ \qquad\qquad \text{if } 2h_L(l) \leq C < 2h_L(l-1), \\ \frac{C^p}{2}, \\ \qquad\qquad \text{if } C \leq (2h_L(L) \wedge d_{\min}). \end{cases}$$

For $\alpha = 2p$ we put $L = \lfloor \frac{1}{\alpha} \log_2(|\mathcal{X}|) \rfloor$ and get

$$\mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C) \leq \begin{cases} \frac{C^p}{2} - 2^{p-1} \left(2 - |\mathcal{X}|^{-1/\alpha}\right)^p \mathrm{diam}(\mathcal{X})^p \\ \quad + 2^{3p-1} \mathrm{diam}(\mathcal{X})^p \left(2^{-p} + A^{1/2} \frac{1}{\alpha} \log_2(|\mathcal{X}|)\right), \\ \qquad \text{if } C \geq 2h_L(0), \\ 2^{3p-1} \mathrm{diam}(\mathcal{X})^p \left(2^{-p} + A^{1/2} \left(\frac{1}{\alpha} \log_2(|\mathcal{X}|) - l\right)\right), \\ \qquad \text{if } 2h_L(l) \leq C < 2h_L(l-1), \\ \frac{C^p}{2}, \\ \qquad \text{if } C \leq (2h_L(L) \wedge d_{\min}). \end{cases}$$

For $\alpha > 2p$ and $L = \lfloor \frac{1}{\alpha} \log_2(|\mathcal{X}|) \rfloor$, it holds

$$\mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C) \leq \begin{cases} \frac{C^p}{2} - 2^{p-1} \left(2 - |\mathcal{X}|^{-1/\alpha}\right)^p \mathrm{diam}(\mathcal{X})^p \\ \quad + 2^{3p-1} \mathrm{diam}(\mathcal{X})^p |\mathcal{X}|^{1/2-p/\alpha} \left(2^{-p} + A^{1/2} \frac{2^{\alpha/2-2p}}{2^{\alpha/2-p}-1}\right), \\ \qquad \text{if } C \geq 2h_L(0), \\ 2^{3p-1} \mathrm{diam}(\mathcal{X})^p \left(2^{-p} |\mathcal{X}|^{1/2-p/\alpha} \right. \\ \quad \left. + A^{1/2} \frac{2^{\alpha/2-p}}{2^{\alpha/2-p}-1} \left(|\mathcal{X}|^{1/2-p/\alpha} - 2^{(\alpha/2-p)(l-1)}\right)\right), \\ \qquad \text{if } 2h_L(l) \leq C < 2h_L(l-1), \\ \frac{C^p}{2}, \\ \qquad \text{if } C \leq (2h_L(L) \wedge d_{\min}). \end{cases}$$

These upper bounds depend on the penalty C as well as the parameters α and A which dictates the behavior of the covering number $\mathcal{N}(\mathcal{X}, \varepsilon)$ of $\mathcal{X}$ at difference scales ε.

If $\alpha < 2p$, then there is no dependence on $|\mathcal{X}|$ and the convergence of approximation error of the empirical measure is independent of its support size. If $\alpha = 2p$, then $|\mathcal{X}|$ enters through a logarithmic term. If $\alpha > 2p$, then the dependence becomes polynomial in $|\mathcal{X}|$. These phase transitions for the dependence on the support size align with those for empirical (balanced) OT for an arbitrary finite subset of $\mathbb{R}^D$ (Sommerfeld et al., 2019), corresponding to $\alpha = D$. Our results also conceptually align with rate results for the empirical Wasserstein distance (Fournier and Guillin, 2015; Weed and Bach, 2019), namely that for $\alpha < 2p$ parametric convergence rates in the sample size and independent of the domain manifest. Moreover, for $\alpha > 2p$ some polynomial dependency in $|\mathcal{X}|$ has to manifest, since otherwise it would imply generic parametric convergence rates in the sample size for the empirical Wasserstein distance for $\alpha > 2p$, which would contradict well-known lower bounds (Singh and Póczos, 2018; Weed and Bach, 2019). We conjecture the dependency in $|\mathcal{X}|$ for the different regimes to be sharp.

A novelty for the UOT setting is the additional dependency on the different scales of C. This is explained by the previously discussed control of C on the maximal distance at which

transport occurs in an optimal plan. The height function is again used to specify the scale induced by a particular choice of the parameter C. Notably, the dependence on $|\mathcal{X}|$ does not change on most scales of C. There is an exception, however, for sufficiently small values of C, where the (p, C)-KRD is equal to a scaled total variation distance. Thus, these bounds are completely oblivious to the geometry of $\mathcal{X}$ in $\mathcal{Y}$, though they scale as $|\mathcal{X}|^{\frac{1}{2}}$ which is the same rate we obtain for the TV bound (recall Theorem 4). As a final observation, we note that for $C > 2h_L(0)$, these bounds essentially recover analogue bounds for empirical optimal transport (Sommerfeld et al., 2019). However, for the (p, C)-KRD the bounds include an additional summand based on the estimation error for the measure's total mass intensity.

Remark 10 *Assumption* (19) *on $\mathcal{X}$ can be relaxed in terms of ε. Namely, for $C \leq 2h_L(L) \wedge d_{\min}$ the assumption is not required because in this case the UOT cost coincides with the TV distance and the complexity of $\mathcal{X}$ does not play a role. For $C \geq 2h_L(0)$, identical upper bounds in $t > 0$ remain valid if* (19) *is satisfied for $\varepsilon > t^{-1/(2\vee\alpha)}$. This reflects a multiscale behavior of the empirical plug-in estimator previously observed for empirical OT (Weed and Bach, 2019): if the finitely supported measure is concentrated on an ε-fattened low-dimensional domain, then for small t the constant from the low-dimensional setting will manifest, but for $t \to \infty$ the constant degrades due to the fattening.*

3. Empirical Unbalanced Optimal Transport Plans

In this section we derive novel quantitative convergence statements for empirical UOT plans. We rely on a novel stability result for the balanced setting, which we also include as Theorem 14 as it might be interesting on its own; the proof is stated in Appendix D.3. The rest of the omitted proofs can be found in Appendix D.4.

To set notation, we denote the collection of UOT plans between measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ for parameters $p \geq 1$ and $C \geq 0$ by

$$\overline{\mathbf{P}}^*_{p,C}(\mu, \nu) \coloneqq \operatorname*{argmin}_{\pi \in \Pi_{\leq}(\mu,\nu)} \sum_{x,x' \in \mathcal{X}} d^p(x, x')\pi(x, x') + C^p \left(\frac{\mathbb{M}(\mu) + \mathbb{M}(\nu)}{2} - \mathbb{M}(\pi) \right). \tag{20}$$

Moreover, we additionally define the domain $\mathcal{D}(C) \coloneqq \{(x, x') \in \mathcal{X} \times \mathcal{X} | d(x, x') < C\}$ and consider the set of restricted UOT plans

$$\mathbf{P}^*_{p,C}(\mu, \nu) \coloneqq \left\{ \pi \cdot \mathbb{1}(\cdot \in \mathcal{D}(C)) \,\middle|\, \pi \in \overline{\mathbf{P}}^*_{p,C}(\mu, \nu) \right\}. \tag{21}$$

Lemma 36 in Appendix D.4 shows that $\mathbf{P}^*_{p,C}(\mu, \nu)$ is a subset of $\overline{\mathbf{P}}^*_{p,C}(\mu, \nu)$, and that every sub-coupling in $\Pi_{\leq}(\mu, \nu)$ which is the sum of an element in $\overline{\mathbf{P}}^*_{p,C}(\mu, \nu)$ and a non-negative measure supported on $\{(x, x') \in \mathcal{X} \times \mathcal{X} \,|\, d(x, x') = C\}$ is contained in $\overline{\mathbf{P}}^*_{p,C}(\mu, \nu)$. Based on this insight, we restrict our attention to the set of restricted UOT plans for empirical and population measures,

$$\mathbf{P}^*_{p,C}(\widehat{\mu}_{t,s}, \widehat{\nu}_{t,s}) \quad \text{and} \quad \mathbf{P}^*_{p,C}(\mu, \nu),$$

and quantify the accuracy in estimating the latter in terms of the former via the Hausdorff distance induced by the total variation norm. Crucial for our analysis is the following novel stability bound of the UOT plans.

Theorem 11 (Stability bound for UOT plans) *Let $(\mathcal{X}, d)$ be a finite metric space and $\mu^1, \mu^2, \nu^1, \nu^2 \in \mathcal{M}_+(\mathcal{X})$. Then, for any $p \geq 1$ and $C \geq 0$ it holds that*

$$\begin{aligned}&\mathcal{H}_{\mathrm{TV}}\left(\mathbf{P}^*_{p,C}(\mu^1, \nu^1), \mathbf{P}^*_{p,C}(\mu^2, \nu^2)\right)\\ &\leq 4\left(|\mathcal{X}| + 1\right)\left(\mathrm{TV}(\mu^1, \mu^2) + |\mathbb{M}(\mu^1) - \mathbb{M}(\mu^2)| + \mathrm{TV}(\nu^1, \nu^2) + |\mathbb{M}(\nu^1) - \mathbb{M}(\nu^2)|\right),\end{aligned}$$

where $\mathcal{H}_{\mathrm{TV}}$ denotes the Hausdorff distance induced by the total variation norm.

The proof is based on relating the collection of (restricted) UOT plans to associated OT plans for suitably augmented measures and a stability bound for OT plans stated in Theorem 14. The latter follows from a general stability bound for optimal solutions of linear program theory based on Li (1994). A remarkable property of the above stability bound is that it does not depend on parameters p or C. In fact, the assertion also remains valid for any other cost function besides of $c = d^p$, and might be of independent interest. This stability bound asserts at the main result of this section.

Theorem 12 (Expected deviation of empirical UOT plans) *Let $(\mathcal{X}, d)$ be a finite metric space and $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$. Let $\widehat{\mu}_{t,s}, \widehat{\nu}_{t,s}$ be the estimator from (8). Then, for any $p \geq 1$ and $C \geq 0$ it follows that*

$$\begin{aligned}&\mathbb{E}\left[\sup_{p\geq 1, C>0} \mathcal{H}_{\mathrm{TV}}\left(\mathbf{P}^*_{p,C}(\widehat{\mu}_{t,s}, \widehat{\nu}_{t,s}), \mathbf{P}^*_{p,C}(\mu, \nu)\right)\right]\\ &\leq 4\left(|\mathcal{X}| + 1\right)\Bigg(2(1-s)\mathbb{M}(\mu+\nu) + \frac{s}{\sqrt{t}}\sum_{x\in\mathcal{X}}\left[\sqrt{\mu(x)} + \sqrt{\nu(x)}\right]\\ &\qquad + \sqrt{\frac{1}{st}\mathbb{M}(\mu+\nu) + \frac{1-s}{s}\sum_{x\in\mathcal{X}}(\mu(x)^2 + \nu(x)^2)}\Bigg).\end{aligned}$$

Proof The assertion follows by combining the stability bound from Theorem 11 above with the bound on the total variation distance between the estimators $\widehat{\mu}_{t,s}, \widehat{\nu}_{t,s}$ and μ, ν (Theorem 4) and on the absolute deviation between their masses (31). ■

Remark 13 (Computation) *Thanks to the connection between UOT and balanced OT (see Appendix C), various polynomial time procedures exist to compute the UOT plan between plug-in estimators $\widehat{\mu}_{t,s}$ and $\widehat{\nu}_{t,s}$. For exact computation, the auction algorithm (Bertsimas and Tsitsiklis, 1997) or the network simplex flow algorithm (Luenberger et al., 1984; Bonneel et al., 2011) can be used which admit a computational complexity of order $\mathcal{O}(N^3 \log(N))$ where N is the number of support points. Moreover, to approximate the*

UOT plan up to some precision $\varepsilon > 0$, the Sinkhorn algorithm for the entropy regularized OT problem (Cuturi, 2013; Peyré and Cuturi, 2019) provides a computationally effective method which scales with order $\mathcal{O}(N^2 \log(N)/\varepsilon^2)$, see Altschuler et al. (2017); Dvurechensky et al. (2018) as well as Weed (2018) for an explicit analysis on the difference between unregularized and entropy regularized OT cost. Notably, incorporating additionally the structure of the UOT plan, any such algorithm can be improved by restricting the measures on the domain $\mathcal{D}(C)$.

For our stability bound of the UOT plan we establish a novel stability bound for the vanilla OT plan between measures with equal mass which might be of independent interest. To formalize these results, we employ the following notation. Let $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ be measures with identical mass $\mathbb{M}(\mu) = \mathbb{M}(\nu)$, let $c\colon \mathcal{X} \times \mathcal{X} \to [0, \infty)$ be a cost function and denote the respective collection of OT plans as

$$\widetilde{\mathbf{P}}_c^*(\mu, \nu) \coloneqq \operatorname*{argmin}_{\pi \in \Pi_=(\mu,\nu)} \sum_{x,x' \in \mathcal{X}} c(x, x')\pi(x, x'). \tag{22}$$

Note that the assumption of identical masses ensures that the collection of transport plans is always non-empty and compact, hence $\widetilde{\mathbf{P}}_c^*(\mu, \nu)$ is always non-empty.

Theorem 14 (Stability bound for balanced OT plans) *Let $\mathcal{X}$ be a finite discrete space and $\mu_1, \nu_1, \mu_2, \nu_2 \in \mathcal{M}_+(\mathcal{X})$ such that $\mathbb{M}(\mu_1) = \mathbb{M}(\nu_1)$ and $\mathbb{M}(\mu_2) = \mathbb{M}(\nu_2)$. Then, for every cost function $c\colon \mathcal{X} \times \mathcal{X} \to [0, \infty)$ it holds that*

$$\mathcal{H}_{\mathrm{TV}}\left(\widetilde{\mathbf{P}}_c^*(\mu_1, \nu_1), \widetilde{\mathbf{P}}_c^*(\mu_2, \nu_2)\right) \leq 4\,|\mathcal{X}|\,(\mathrm{TV}(\mu_1, \mu_2) + \mathrm{TV}(\nu_1, \nu_2))\,.$$

The proof is based on linear program theory and relies on a stability bound by Li (1994).

This quantitative error bound represents, to the best of our knowledge, the first of its kind for finite discrete domains. Moreover, thanks to the generality of our stability bound, we are not limited to the setting $p = 2$, which has been the main subject of analysis in Euclidean settings for continuous settings (Gigli, 2011; Deb et al., 2021a; Hütter and Rigollet, 2021; Delalande and Merigot, 2023; Manole et al., 2024a) as well as semi-discrete settings (Bansil and Kitagawa, 2022; Divol et al., 2024). Notably, these former results are concerned with the OT map, whereas our Theorem 14 is concerned with the OT plan. It remains open whether similar deviation bound for the OT plan can also be established for continuous settings. Since the total variation norm metrizes strong convergence, we believe that similar results are will not be valid for empirical plug-in estimators on continuous domains but can be achieved when choosing a weaker loss and smooth plug-in estimators. We leave this for future research.

Remark 15 *Let us address some aspects of our stability bound from Theorem 14.*

1. *The key insight of Theorem 14 is the fact that they behave fairly stable under small perturbation of the marginal measures with respect to TV-norm. Remarkably, the underlying cost function does not affect the upper bound, which may seem surprising given that the OT plan crucially depends on the cost function (see, e.g., Villani (2008)).*

2. *For $\mu^1 = \mu^2 = \delta_x$ and arbitrary measures $\nu^1, \nu^2 \in \mathcal{M}_+(\mathcal{X})$ it follows that $\widetilde{\mathbf{P}}_c^*(\mu_i, \nu_i) = \{\mu_i \otimes \nu_i\}$. Therefore,*

$$\mathcal{H}_{\mathrm{TV}}\left(\widetilde{\mathbf{P}}_c^*(\mu_1, \nu_1), \widetilde{\mathbf{P}}_c^*(\mu_2, \nu_2)\right) = \mathrm{TV}(\mu_1 \otimes \nu_1, \mu_2 \otimes \nu_2) = \mathrm{TV}(\nu_1, \nu_2),$$

and thus the dependency in Theorem 14 with respect to the marginal measures is sharp.

3. *The upper bound only scales linearly in the number of elements of the domain $\mathcal{X}$. If all measures $\mu^1, \nu^1, \mu^2, \nu^2$ are concentrated on a subdomain $\mathcal{X}' \subseteq \mathcal{X}$, the constant could be replaced by $|\mathcal{X}'|$. Obtaining the sharp dependency in the number of support points remains an interesting aspect for future work.*

4. *Based on distributional limits for the empirical OT plan between discrete measures (Klatt et al., 2022; Liu et al., 2023) it follows that the deviation bound in Theorem 12 is sharp in s and t up to multiplicative constants which depend on $\mathcal{X}$, μ and ν.*

5. *When perturbing the cost function a similar stability bound for OT plans can generally not be expected. As an example, take identical probability measures $\mu = \nu$ which are not concentrated on a single point and consider $c_a(x, y) = a \cdot d(x, y)$ as the cost function where $a \in [0, \infty)$ and d is a metric on $\mathcal{X}$. Then, for $a > 0$ it follows that $\widetilde{\mathbf{P}}_{c_a}^*(\mu, \nu) = \{(\mathrm{Id}, \mathrm{Id})_{\#}\mu\}$ whereas for $a = 0$ it holds $\widetilde{\mathbf{P}}_{c_0}^*(\mu, \nu) = \Pi(\mu, \nu)$. This confirms that the set of OT plans is discontinuous with respect to the cost function since*

$$\lim_{a \searrow 0} \mathcal{H}_{\mathrm{TV}}\left(\widetilde{\mathbf{P}}_{c_a}^*(\mu, \nu), \widetilde{\mathbf{P}}_{c_0}^*(\mu, \nu)\right) > 0 \quad \text{while} \quad \lim_{a \searrow 0} \|c_a - c_0\|_\infty = 0.$$

4. Empirical Kantorovich-Rubinstein Barycenters

Consider measures $\mu^1, \ldots, \mu^J \in \mathcal{M}_+(\mathcal{X})$ which we replace by $\widehat{\mu}_{t_1,s_1}^1, \ldots, \widehat{\mu}_{t_J,s_J}^J \in \mathcal{M}_+(\mathcal{X})$ as defined in (8). We again focus on the Poisson model and treat the remaining two models in Appendices A and refsec:ber. The previous upper bound on the Kantorovich-Rubinstein distance in Theorem 5 between a measure and its empirical version enables a bound on the mean absolute deviation of (p, C)-barycenters in terms of their p-Fréchet functional $F_{p,C}(\mu) = \frac{1}{J}\sum_{i=1}^J \mathrm{KR}_{p,C}^p(\mu^i, \mu)$ from (3). We denote

$$\mu^\star \in \operatorname*{argmin}_{\mu \in \mathcal{M}_+(\mathcal{Y})} \frac{1}{J}\sum_{i=1}^J \mathrm{KR}_{p,C}^p(\mu^i, \mu), \quad \widehat{\mu}^\star \in \operatorname*{argmin}_{\mu \in \mathcal{M}_+(\mathcal{Y})} \frac{1}{J}\sum_{i=1}^J \mathrm{KR}_{p,C}^p(\widehat{\mu}_{t_i,s_i}^i, \mu)$$

and measure the accuracy of approximation of $\mu^\star$ by $\widehat{\mu}^\star$ in terms of their mean absolute p-Fréchet deviation. The omitted proofs for this section are detailed in Appendix D.5.

Theorem 16 (Expected deviation of Fréchet error) *Let $\mu^1, \ldots, \mu^J \in \mathcal{M}_+(\mathcal{X})$ and denote $\mathcal{X}_i = \mathrm{supp}(\mu^i)$ for $i = 1, \ldots, J$. Consider random estimators $\widehat{\mu}_{t_1,s_1}^1, \ldots, \widehat{\mu}_{t_J,s_J}^J \in \mathcal{M}_+(\mathcal{X})$ derived from (8). Then,*

$$\mathbb{E}\left[|F_{p,C}(\widehat{\mu}^\star) - F_{p,C}(\mu^\star)|\right] \leq \frac{2pC^{p-1}}{J}\sum_{i=1}^J \mathcal{E}_{1,\mathcal{X}_i,\mu^i}^{\mathrm{Poi}}(C)\phi(t_i, s_i),$$

where ϕ is given by

$$\phi(t,s) = \begin{cases} \left(2(1-s)\mathbb{M}(\mu) + \frac{s}{\sqrt{t}} \sum_{x\in\mathcal{X}} \sqrt{\mu(x)}\right), & \text{if } C \leq \min_{x\neq x'} d(x,x'), \\ \left(\frac{1}{st}\mathbb{M}(\mu) + \frac{1-s}{s} \sum_{x\in\mathcal{X}} \mu(x)^2\right)^{\frac{1}{2}}, & \text{else.} \end{cases}$$

A more elaborate statement gives control over the set of empirical (p, C)-barycenters itself. This involves a related linear program that is presented in detail in Appendix C.

Theorem 17 (Expected deviation of barycenters) *Let $\mu^1, \dots, \mu^J \in \mathcal{M}_+(\mathcal{X})$ and denote $\mathcal{X}_i = \operatorname{supp}(\mu^i)$ for $i = 1, \dots, J$. Consider random estimators $\widehat{\mu}^1_{t_1,s_1}, \dots, \widehat{\mu}^J_{t_J,s_J} \in \mathcal{M}_+(\mathcal{X})$ derived from (8). Let $\mathbf{B}^\star$ be the set of (p, C)-barycenters of $\mu^1, \dots, \mu^J$ and $\widehat{\mathbf{B}}^\star$ the set of (p, C)-barycenters of $\widehat{\mu}^1_{t_1,s_1}, \dots, \widehat{\mu}^J_{t_J,s_J}$. Then, for $p \geq 1$ it holds that*

$$\mathbb{E}\left[\sup_{\widehat{\mu}^\star\in\widehat{\mathbf{B}}^\star} \inf_{\mu^\star\in\mathbf{B}^\star} \mathrm{KR}^p_{p,C}(\widehat{\mu}^\star, \mu^\star)\right] \leq \frac{pC^{p-1}}{V_P J} \sum_{i=1}^{J} \mathcal{E}^{\mathrm{Poi}}_{1,\mathcal{X}_i,\mu^i}(C)\phi(t_i, s_i),$$

where ϕ is defined as in Theorem 16. The constant V_P is strictly positive and given by

$$V_P \coloneqq V_P(\mu^1, \dots, \mu^J) := (J+1)\operatorname{diam}(\mathcal{X})^{-p} \min_{v\in V\setminus V^\star} \frac{c^T v - f^*}{d_1(v, \mathcal{M})},$$

where V is the set of feasible vertices from the linear program in Appendix C, $V^\star$ is the subset of optimal vertices, c is the cost vector of the program, $f^\star$ is the optimal value, $\mathcal{M}$ is the set of minimizers of the linear program (29) and $d_1(x, \mathcal{M}) = \inf_{y\in\mathcal{M}} \|x - y\|_1$.

Remark 18 *Let us comment on our convergence result for barycenters.*

1. *Theorem 17 implies that every empirical barycenter tends for $\min_{i\in\{1,\dots,J\}} t_i \to \infty$ and $\min_{i\in\{1,\dots,J\}} s_i \to 1$ with parametric rates towards a suitable population barycenter. This is practically relevant because for discrete domains, non-uniqueness of empirical and population barycenters generally cannot be ruled out. Insofar, our result guarantees that every estimator admits satisfactory convergence properties.*

2. *According to Hundrieser et al. (2024a), it is known that any barycenters estimator suffers nearby the regime of non-unique barycenters from slow convergence rates or underlying constants which blow up. Based on this insight, the presence of the term V_P, which can be interpreted as a sub-optimality gap, likely cannot be improved significantly without imposing additional conditions on the population measures $\mu_1, \dots, \mu_J$.*

3. *As an extension of Theorem 17 we deem it worthwhile to extend the analysis to the Hausdorff distance $\mathcal{H}_{\mathrm{KR}_{p,C}}$ induced by the (p, C)-KRD, i.e., to establish convergence statements for empirical barycenters $\widehat{\mathbf{B}}^\star$ to population barycenters $\mathbf{B}^\star$,*

$$\mathcal{H}^p_{\mathrm{KR}_{p,C}}\left(\widehat{\mathbf{B}}^\star, \mathbf{B}^\star\right) = \max\left(\sup_{\widehat{\mu}^\star\in\widehat{\mathbf{B}}^\star} \inf_{\mu^\star\in\mathbf{B}^\star} \mathrm{KR}^p_{p,C}(\mu^\star, \widehat{\mu}^\star), \sup_{\mu^\star\in\mathbf{B}^\star} \inf_{\widehat{\mu}^\star\in\widehat{\mathbf{B}}^\star} \mathrm{KR}^p_{p,C}(\mu^\star, \widehat{\mu}^\star)\right).$$

Such a convergence statement in the Hausdorff distance would assert, in addition to point 1., that for every population barycenter there exists an empirical barycenter which approximates it. Arguing as in Appendix D.5, one can see that

$$\widehat{V}_P \sup_{\mu^\star \in \mathbf{B}^\star} \inf_{\widehat{\mu}^\star \in \widehat{\mathbf{B}}^\star} \mathrm{KR}^p_{p,C}(\mu^\star, \widehat{\mu}^\star) \leq |\widehat{F}_{p,C}(\widehat{\mu}^\star) - \widehat{F}_{p,C}(\mu^\star)|,$$

where $\widehat{V}_P$ is defined analogously to V_P but with each μ^i replaced by the estimator $\widehat{\mu}^i_{t_i,s_i}$. Controlling $\widehat{V}_P$ to infer quantitative convergence results, however, seems rather challenging, and we leave a refined analysis of the underlying constant for future research.

Remark 19 *For $J = 1$ and any $p \geq 1, C > 0$ the (p, C)-barycenter of μ^1 is just μ^1. Thus, the optimal value of the Fréchet functional is zero, and it holds*

$$F_{p,C}(\widehat{\mu}^*) - F_{p,C}(\mu^*) = \mathrm{KR}^p_{p,C}(\mu^1, \widehat{\mu}^1_{t,s}).$$

Consequently, it also holds

$$\sup_{\widehat{\mu}^\star \in \widehat{\mathbf{B}}^\star} \inf_{\mu^\star \in \mathbf{B}^\star} \mathrm{KR}^p_{p,C}(\mu^*, \widehat{\mu}^*) = \mathrm{KR}^p_{p,C}(\mu^1, \widehat{\mu}^1_{t,s}).$$

Thus, the rate for the convergence of the (p, C)-barycenter of the empirical measures, can in general not be faster than the convergence rate of a single estimator. In particular, the rates in t in Theorem 16 and Theorem 17 are sharp.

Remark 20 *While this work is only concerned with plug-in estimators for (p, C)-barycenters between general measures, there are also various other notions of barycenters between general measures Chizat et al. (2018a); Friesecke et al. (2021). Exploring the statistical properties of these alternative barycenters presents an interesting venue for future research.*

5. Application: Randomized Computation with Statistical Guarantees

In this section we discuss how our derived bounds enable randomized computations of UOT quantities with statistical guarantees. If the size of the population measures is computationally infeasible, then empirical versions of these measures can be used as a proxy for the population distance, plan and barycenter. Our bounds allow tuning the problem size (hence computational time) against the accuracy of the approximation. While all three models allow this approximation approach, we exemplify this for the multinomial model. In this resampling scenario, the assumption of known total intensities amounts to have access to the full data set. Here, the sample size N provides a strict upper bound on the computational complexity of a given approximation. This is though not the case for the other two models, where such bounds can only be obtained by involving subsampling [10] scheme instead of a resampling one. We note however that subsampling approach is typically outperformed by the resampling one.

10. Here, subsampling refers to sampling without replacement, while resampling refers to sampling with replacement.

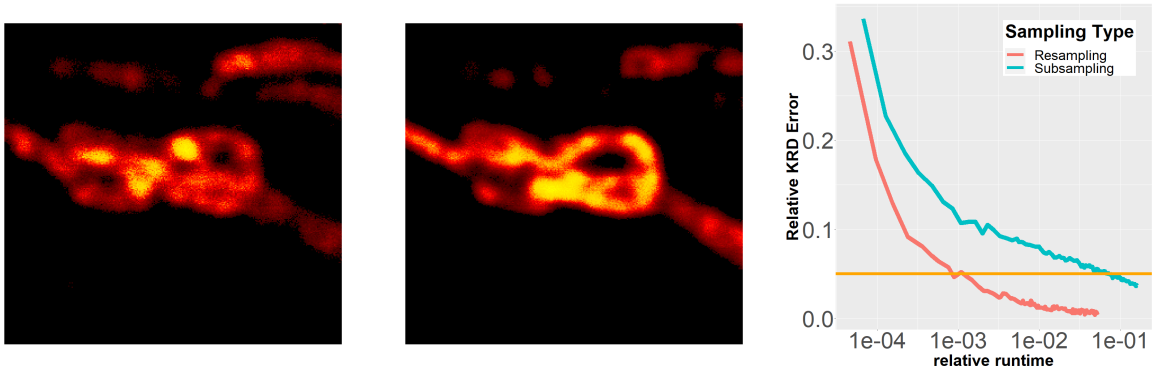


Figure 5: **Left:** An excerpt of size 300×300 from the STED microscopy data of adult Human Dermal Fibroblasts labelled at MIC60 (a mitochondrial inner membrane complex, see Tameling et al. (2021)). **Center:** The same type of data labelled at TOM20 (translocase of the outer mitochondrial membrane), see Tameling et al. (2021). **Right:** The expected relative $(2, 0.1)$-KRD error curves obtained for the two images on the left and in the center from re- and subsampling for specified runtimes (computed with the CRAN package *WS-Geometry*). The orange line indicates an error level of 5%. The respective sample sizes are between 100 and 9000 for both approaches, while the original images have about 23000 non-zero pixels.

Indeed, for the subsampling scheme we replace the i.i.d. samples $X_1, \dots, X_N$ from μ by ones drawn without replacement. A natural choice of estimator is

$$\widetilde{\mu}_N = \frac{\mathbb{M}(\mu)}{\sum_{i=1}^N \mu(X_i)} \sum_{i=1}^N \mu(X_i)\delta_{X_i}, \tag{23}$$

where the mass at each drawn location $x \in \mathcal{X}$ is proportional to the mass of the population measure at x and the total mass intensity is rescaled to the known, true total intensity. This estimator is, due to the sampling without replacement, guaranteed to have N support points, which yields close control on the required runtime for a given approximation. In recent years, this approach has become popular within the machine learning community where it is referred to as mini-batch OT (Fatras et al., 2021; Nguyen et al., 2022). An illustration of the potential runtime advantages using the suggested randomized methods is displayed in Figure 5. For this example, the resampling approach provides an expected relative KRD error of about 5% while achieving a speedup of about a factor of 1000 compared to the original runtime, while the subsampling approach requires nearly 10% of the original runtime to achieve the same accuracy. A more detailed comparison of the empirical performance of the re- and subsampling model is found in Section 6.4. Though, we note that, in the considered data examples, the resampling approach consistently performed better than the subsampling one. We also study the convergence properties of the empirical measure and barycenter with respect to the (p, C)-KRD for all three described models in extended simulation studies on a wide range of synthetic datasets, again further in Section 6.

6. Simulations

In this section we investigate empirically the decay in the expected error for the Poisson model for measures within $\mathcal{X} \subset [0,1]^2$. For the (p,C)-KRD we consider two measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ and the *relative* (p,C)*-KRD error*[11]

$$\mathbb{E}\left[\left|\frac{\mathrm{KR}_{p,C}(\widehat{\mu}_{t,s}, \widehat{\nu}_{t,s}) - \mathrm{KR}_{p,C}(\mu,\nu)}{\mathrm{KR}_{p,C}(\mu,\nu)}\right|\right]. \tag{24}$$

For the setting of barycenters we consider the *relative* (p,C)*-Fréchet error*[11]

$$\mathbb{E}\left[\frac{F_{p,C}(\widehat{\mu}^*) - F_{p,C}(\mu^*)}{F_{p,C}(\mu^*)}\right]. \tag{25}$$

In both cases, the relative error allows for easier comparisons between models than the absolute error. In particular, since $\mathrm{KR}^p_{p,C}(\mu,\nu) = \frac{C^p}{2}\mathrm{TV}(\mu,\nu)$ (Heinemann et al., 2023, Theorem 2) for $C \searrow 0$ and $\mu \neq \nu$, it follows that the relative (p,C)-KRD error enables an easier comparison of the estimation error among different choices of C. Additionally, for the (p,C)-barycenter the relative (p,C)-Fréchet error in (25) is readily available from simulations, while numerically considering the quantity in Theorem 17 is difficult, as it requires all optimal solutions instead of a single one. All computations of the KRD and the (p,C)-barycenter in this section are performed using the methods available in the CRAN package *WSGeometry*.

6.1 Synthetic Datasets

We consider multiple types of measures for our simulations. Below we describe four types; Appendix E.1 contains four additional types and the respective Poisson simulations are provided in Appendix E.2. Analogous simulations for Multinomial Sampling and Bernoulli Sampling are detailed in Appendices E.3 and E.4, respectively.

Let us fix some notation. Let $J \in \mathbb{N}$ be the number of measures generated. Let $U[0,1]^2$ denote the uniform distribution and let $\mathrm{Poi}(\lambda)$ denote a Poisson distribution with intensity λ. In all settings considered below, the measures are of the form

$$\mu^i = \sum_{k=1}^{K_i} w_k^i \delta_{l_k^i}$$

for some weights w_k^i, locations l_k^i and $K_i \in \mathbb{N}$. If $K_i = K_j$ for all $i,j = 1,\dots,J$, then we omit the index and denote the number of points by K. Note that all measures have been constructed to have their support included in $[0,1]^2$.

Nested Ellipses (NE), see Figure 6 (a)

Let $G_1,\dots,G_J \sim U\{1,2,3,4,5\}$ and let $K_i = MG_i$ for $M \in \mathbb{N}$. Set all w_k^i equal to 1 for each $1 \leq k \leq K_i$ for $i = 1,\dots,J$. Let $t_1,\dots,t_M$ be a discretization of $[0,2\pi]$. Let

11. We define $0/0 := 0$.

$U_1^i, \ldots, U_K^i, V_1^i, \ldots, V_K^i \sim U[0.2, 1]$. For $1 \leq i \leq J$, set

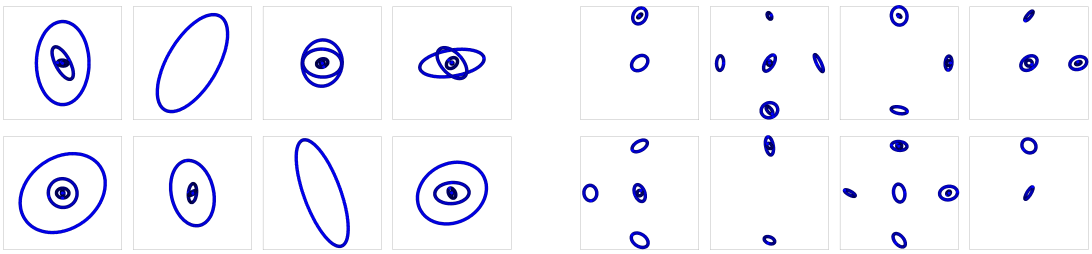
$$l^i_{M(j_i-1)+k} = 0.5(1 + 3^{-j}(U_{M(j_i-1)+k} \sin(t_k), V_{M(j_i-1)+k} \cos(t_k))^T), \quad j_i = 0, \ldots, G_i.$$

Clustered Nested Ellipses (NEC), see Figure 6 (b)

Let $G_1^c, \ldots, G_J^c \sim \text{Poi}(\lambda_c)$ for $c = 1, \ldots, 5$. Take $\lambda_3 = 2$ and set $\lambda_c = 1$ else. Let $K_i = M \sum_{c=1}^5 G_i^c$. Set w_k^i equal to 1 for $1 \leq k \leq K_i$ for $i = 1, \ldots, J$. Let $t_1, \ldots, t_M$ be a discretization of $[0, 2\pi]$. Choose $U_1^i, \ldots, U_K^i, V_1^i, \ldots, V_K^i \sim U[0.2, 1]$. Let $\alpha = (2, 12, 12, 22, 12)^T$ and $\beta = (12, 2, 12, 12, 22)^T$. Set for $c = 1, \ldots, 5$ and $j_i = 0, \ldots, G_i^c$

$$l^i_{M(\sum_{r=1}^{c-1} G_i^r)+M(j_i-1)+k} = \frac{1}{24}((3^{-j} U_{M(j_i-1)+k} \sin(t_k) + \alpha_c, 3^{-j} V_{M(j_i-1)+k} \cos(t_k) + \beta_c))^T,$$

where we use the convention that a sum is zero if its last index is smaller than its first one.



(a) (b)

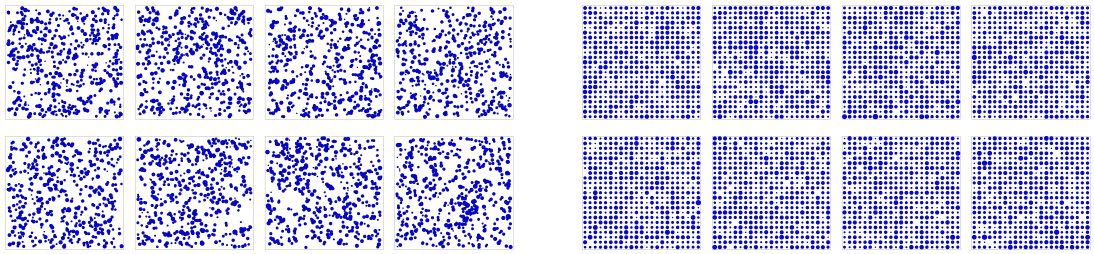
Figure 6: **(a)** An example of $J = 8$ measures from the *NE* dataset with $M = 200$. **(b)** An example of $J = 8$ measures from the *NEC* dataset with $M = 95$.

Poisson Intensities on Uniform Positions (PI), see Figure 7 (a)

Set $K = M$ for $M \in \mathbb{N}$ and let $w_1^i, \ldots, w_K^i \sim \text{Poi}(\lambda)$ and some intensity $\lambda > 0$ and $l_1^i, \ldots, l_K^i \sim U[0, 1]^2$ for $1 \leq i \leq J$.

Poisson Intensities on a Grid (PIG), see Figure 7 (b)

Set $K = M^2$ for $M \in \mathbb{N}$ and let $w_1^i, \ldots, w_{M^2}^i \sim \text{Poi}(\lambda)$ and $l_1^i, \ldots, l_{M^2}^i$ be the locations of equidistant $M \times M$ grid points in $[0, 1]^2$ for $1 \leq i \leq J$.



(a) (b)

Figure 7: **(a)** An example of $J = 8$ measures from the *PI* dataset with $M = 500$ and $\lambda = 5$. **(b)** An example of $J = 8$ measures from the *PIG* dataset with $M = 23$ and $\lambda = 5$.

6.2 Simulation Results for the $(2, C)$-Kantorovich-Rubinstein Distance

In the following, we discuss the results from our simulation studies for the Poisson model for the $(2, C)$-KRD between two measures within one of the eight classes of measures introduced above. For the error of the NE class in Figure 8 the error is decreasing in s and t, but increasing in C. Both of these behaviors are in line with the bound in Theorem 5. The decrease of the error for increasing s and t is immediately clear from our theoretical results. The increase of error for increasing C is based on the fact that in the Poisson model the population total intensities of μ and ν are unknown and have to be estimated from the data. The (p, C)-KRD penalizes mass deviation with a factor scaling with C, so naturally for increasing C, the errors in the estimation of the true difference of masses yields an increase in the expected relative (p, C)-KRD error.

Notably, while the decrease in s and t is similar for the error of the NEC class in Figure 9, the error is no longer increasing in C. Instead, the errors increase from $C = 0.01$ to $C = 0.1$, but then decrease again for $C = 1$. Afterward they increase again at $C = 10$. This difference in behavior is explained by the cluster structure of the measure in NEC. There is still the general trend of increasing error for increasing C, as present in the NE class, but now there is an additional change in behavior based on the fact if transport occurs within clusters or between clusters. From $C = 0.1$ to $C = 1$, we pass the size of the clusters and the distance between the clusters. Thus, $C = 1$ is the first value in our simulation for which inter-cluster transports can occur. This causes a decrease in error, as the impact of the estimation of the total mass intensity of a measure within one cluster is decreased. After this point the usual increase in error for increasing C due to the estimation of the total mass intensity occurs.

For the (p, C)-KRD error in the PI class in Figure 10 this effect is particularly strong. The error increases on average about two orders of magnitude from $C = 0.01$ to $C = 10$. This is explained by the fact that the total mass intensity in this class is significantly larger than for the classes NE and NEC, where each location in the support of the measures has mass one. This also causes an increase of the variance of the mass of the empirical measures at each location, which causes a faster increase of error for increasing C at all scales of C.

Meanwhile, the (p, C)-KRD error in the PIG class in Figure 11 increases from $C = 0.01$ to $C = 0.1$ but does not increase further as C increases. This is likely a consequence of the homogeneous structure of the measures in the PIG class, which cause the UOT plan to transport mass along small scales and since the total mass is well concentrated.

6.3 Simulation Results for the $(\mathbf{2}, \mathbf{C})$-Barycenter

It remains to discuss the results from our simulation studies for the Poisson model for the $(2, C)$-barycenter between sets of measures within one of the classes of measures introduced above. We restrict our analysis to the values of $C = 0.1, 1, 10$, since for $C = 0.01$ the (p, C)-KRD is close to the TV distance. In particular, for all classes except PIG, where all measures share the same support grid, the $(2, C)$-barycenter will be close or identical to the zero measure, since the measures in the other classes are almost surely disjoint. Additionally, if the barycenter of the population measures is the zero measure, any empirical barycenter has mass zero as well. Thus, there is little merit in simulating the barycenters in these cases. For the class PIG the barycenters are essentially TV-barycenters for small C which removes any geometrically interesting features from the barycenter. Finally, for rather small values of C the (p, C)-barycenter computations tend to become numerically unstable due to either involving values close to machine accuracy and UOT plans for these values of C often being close to the zero measure. Hence, empirical simulations of the expected relative Fréchet error would also be less reliable in this regime of values for C. In summary, empirical analysis of the properties of the (p, C)-barycenter for values of C which are several orders of magnitude smaller than the diameter of $\mathcal{Y}$ is inadvisable.

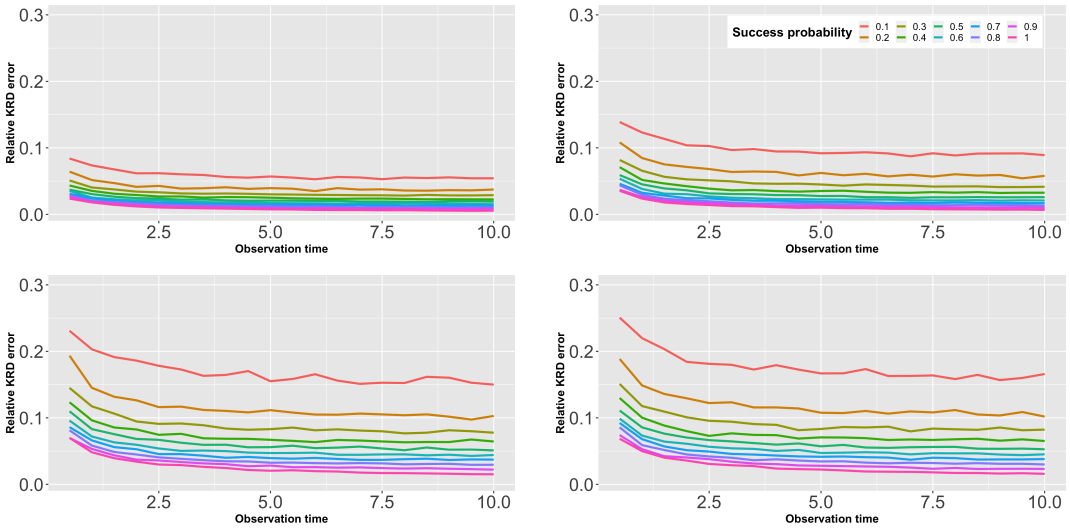


Figure 8: Expected relative $(2, C)$-KRD error for two measures in the Poisson sampling model for the NE class with $M = 100$ and different success probabilities s. For each pair of success probability s and observation time t the expectation is estimated from 1000 independent runs. From top-left to bottom-right we have $C = 0.01, 0.1, 1, 10$, respectively.

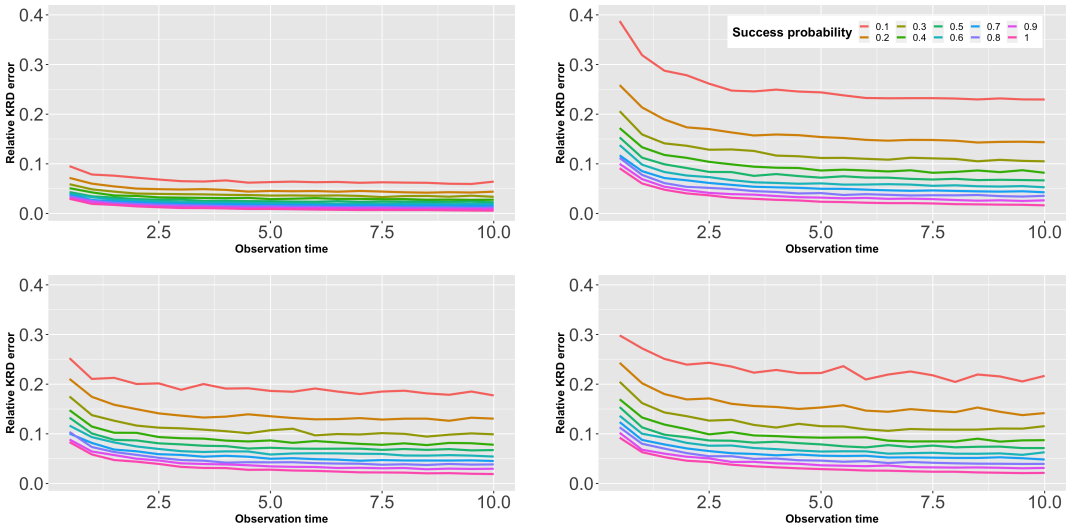


Figure 9: As in Figure 8, but for the NEC class with $M = 75$.

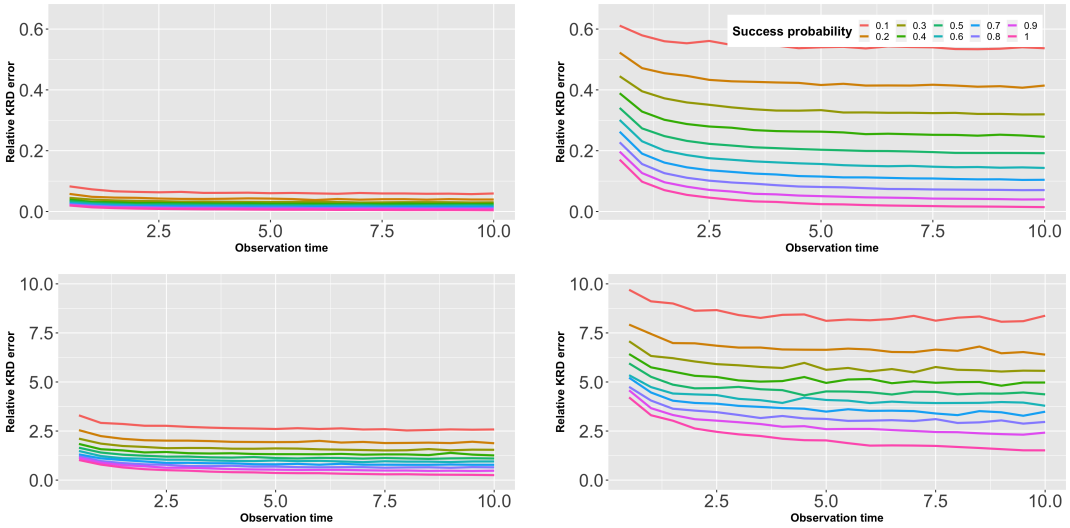


Figure 10: As in Figure 8, but for the PI class with $M = 450$.

For the relative Fréchet errors we observe significant changes in behavior compared to the relative (p, C)-KRD before. Considering the error for the NE class in Figure 12, we note that for $C = 0.1$ the behavior in s and t is different from the empirical (p, C)-KRD. Namely, for fixed s, the error is in general not strictly decreasing in t and vice versa for fixed t, the error is not always strictly decreasing in s. This is an interesting effect arising for small values of the product st. A point $y \in \mathcal{Y}$ can only be a support point of a (p, C)-barycenter if it is in the intersection of at least $J/2$ balls of size $C^p/2$ around support points of different measures (compare the construction of the centroid set in (5)). Now, for small s and t many support points of the population barycenter are not included in the support of the

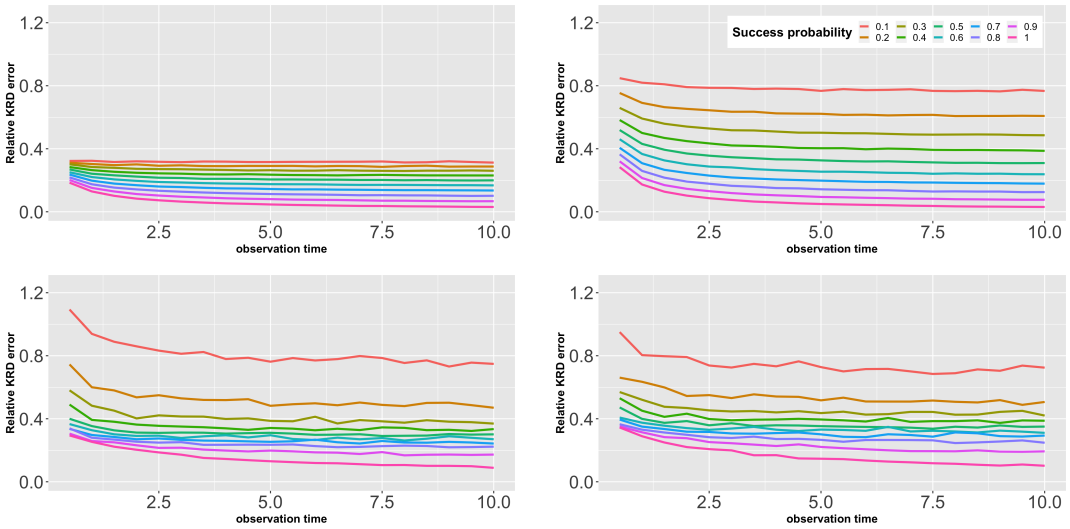


Figure 11: As in Figure 8, but for the PIG class with $M = 22$.

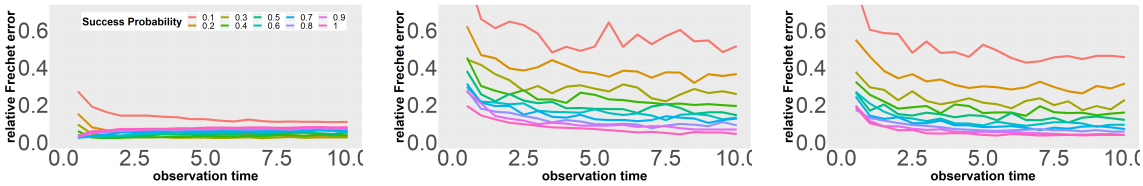


Figure 12: Expected relative Fréchet error for the $(2, C)$-barycenter for $J = 5$ measures from the NE class in the Poisson sampling model with different success probabilities s. For each pair of success probability s and observation time t the expectation is estimated from 100 independent runs. Set $M = 100$. From left to right we have $C = 0.1, 1, 10$, respectively.

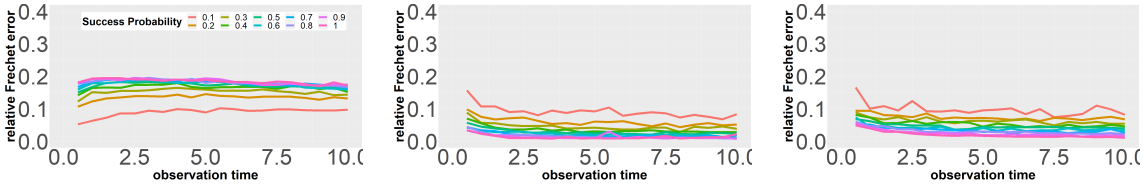


Figure 13: As in Figure 12, but for the NEC class with $M = 75$.

empirical one, since centroid set of the empirical measures is significantly smaller than the population level one. In particular, this can create situations where an increase in s or t on average adds support points to empirical barycenter, which cause the relative error to increase, since placing mass zero at this location, for small C, is actually better than placing a potentially larger mass (since we assumed $(ts)^{-1}$ to be relatively small) at this location. Thus, while asymptotically, the rate in Theorem 16 is optimal, for certain, sufficiently small, values of s, t and C, the behavior of the relative Fréchet error might be counter-intuitive. For $C = 0.1$ and $C = 1$, the errors behave quite similarly to the (p, C)-KRD setting, though

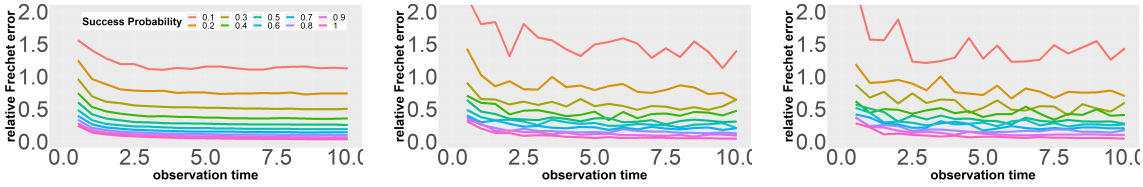


Figure 14: As in Figure 12, but for the PI class with $M = 450$.

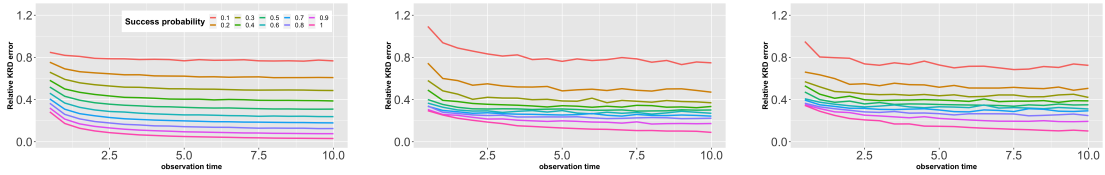


Figure 15: As in Figure 12, but for the PIG class with $M = 22$.

there is essentially no increase in error going from $C = 1$ to $C = 10$. This is explained by two points. First, the location of the (p, C)-barycenter tends to be more centered within the support of the measures (all measures are support on subsets of the unit square), so little transport between the barycenter and the μ^i occurs at a distance larger than one. Second, the key factor for the increasing error for increasing C in the (p, C)-KRD case is the estimation error for the total mass intensities. However, for sufficiently large C the mass of the (p, C)-barycenter is the median of the total masses of the μ^i. Since, this quantity is significantly more stable under estimation than the individual total mass intensities, it is to be expected that the mass estimation has little effect on the relative Fréchet error. For the error of the NEC class in Figure 13 the results are similar to the NE case. We observe similar effects on the dependence of s and t for $C = 0.1$ and for $C = 1$ and $C = 10$, the errors look extremely similar. One notable distinction is the fact that from $C = 0.1$ to $C = 1$ the errors decrease on average. As before for the NEC class in the (p, C)-KRD setting, this can be explained by its cluster structure and $C = 1$ being the first value for which inter-cluster transport becomes possible in an UOT plan. This is therefore also the first value of C which allows the (p, C)-barycenter to have mass between clusters. Finally, for the error of the PI class in Figure 14 the value of C only has a minimal effect on the resulting errors. Notably and contrary to the two prior classes, we do not encounter any additional effects for $C = 0.1$. This is explained by the in general higher mass intensities of the measures in the PI class, which make the previously described effects due to low values of s and t less likely. Additionally, these measures do not possess any geometrical structures in their support, which could impact the behavior on different scales. There is again little increase in error for increasing C, which is in stark contrast to the PI class in the (p, C)-KRD setting, where the error increased by multiple orders of magnitude. This is another strong indicator, that the Fréchet error is significantly more stable under C, due to the stability of total mass intensity of the empirical barycenter opposed to the total mass intensity of the individual measures.

6.4 Real Data Example

In Figure 16 we consider the $(2, 0.1)$-KRD between images which are an excerpt from STED microscopy of adult human dermal fibroblast cells (for the full dataset see Tameling et al. (2021)). The images in the Figures 16(a), (b) and Figures 16(c), (d) are visually similar, as they correspond to measurements taken based on two different markers (one at the inner mitochondrial membrane and one at the outer) in the same cells. The $(2, 0.1)$-KRD captures this fact in the sense, that the pairwise distance between the measures are smallest for these pairs of images. Utilizing UOT on this type of dataset is a potential way of quantifying dissimilarity between the respective measures and extending OT based dissimilarity analysis to measures of unequal total intensity (the total mass intensities in these examples lies roughly between 6200 and 9500).

We further want to use this dataset to illustrate the performance of the randomized computational approach for the (p, C)-KRD based on the multinomial model (recall (6)). The 300×300 images here are specifically chosen such that the true distances can still be computed which allows comparing the expected error of the empirical (p, C)-KRD for given sample sizes on this data set. We compare the results obtained from the resampling approach (i.e., the estimator from (6)) considered in the multinomial model to the subsampling approach (i.e., the estimator from (23)) obtained by sampling without replacement from the measures instead. In these simulations the maximum sample size is about $1/5$ of the support sizes. This corresponds to a runtime of about 2.5% of the original problem size. While it is clear by construction that for sufficiently large sample sizes, subsampling yields a smaller error than the resampling (as the error approaches zero if the sample size approaches the support size), for smaller sample sizes the resampling can have a better performance. It yields a relative error below 5% at less than 10% of the original support size in all considered instances. This approximation can be achieved in around 0.5% of the original runtime. The subsampling approach does not reach this level of accuracy for the considered sample sizes. Thus, these simulations suggest that randomized computations based on the multinomial model allow for high accuracy approximations of the (p, C)-KRD in real data applications at a significantly lower computational cost than the original problem and that for small sample sizes there are scenarios where the resampling approach yields significantly better performance than the subsampling one.

Acknowledgments

S. Hundrieser, F. Heinemann, M. Klatt, and M. Struleva gratefully acknowledge support from the DFG Research Training Group 2088 *Discovering structure in complex data: Statistics meets optimization and inverse problems.* A. Munk gratefully acknowledges support from the DFG CRC 1456 *Mathematics of the Experiment A04, C06*, DFG RU 5381 *Mathematical Statistics in the information age – Statistical efficiency and computational tractability*, and the DFG Cluster of Excellence 2067 MBExC *Multiscale bioimaging–from molecular machines to networks of excitable cells.*

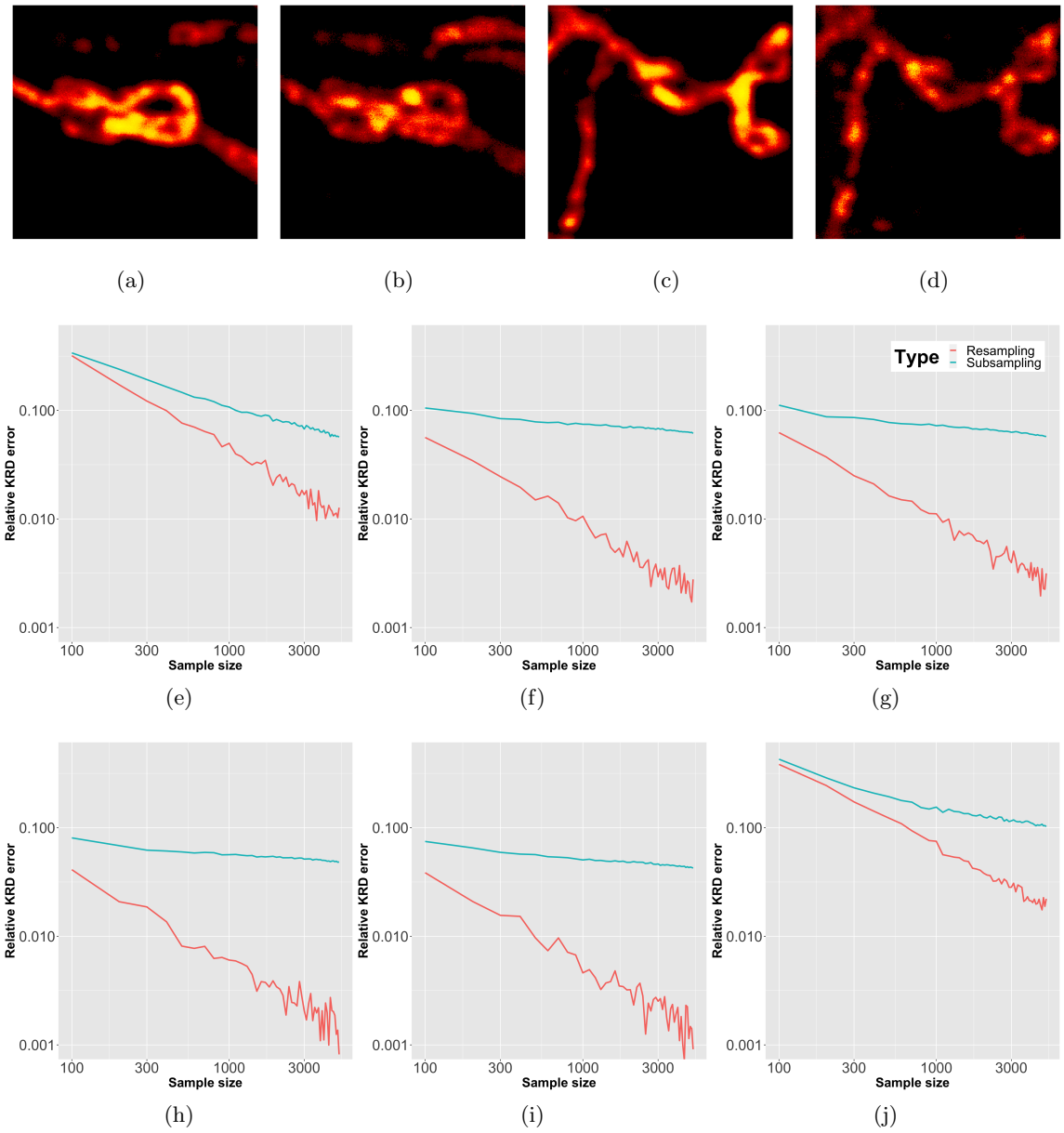


Figure 16: **(a)-(d):** Excerpts of size 300×300 from the STED microscopy data of adult Human Dermal Fibroblasts in Tameling et al. (2021). The images have on average about 25000 non-zero pixels. (a) and (c) have been labelled at MIC60 (a mitochondrial inner membrane complex); (b) and (d) have been labelled at TOM20 (translocase of the outer mitochondrial membrane). **(e)-(j):** Log-log-plots of the relative error for the empirical $(2, 0.1)$-KRD (obtained from resampling and subsampling) between the four filament structures in (a)-(d) considered as measures in $[0, 1]^2$. **(e)** Between (a) and (b). **(f)** Between (a) and (c). **(g)** Between (a) and (d). **(h)** Between (b) and (c). **(i)** Between (b) and (d). **(j)** Between (c) and (d).

Appendix A. Bounds for the Multinomial Model

In this section we provide analogue results to the convergence statement for the expected deviation of the estimator in KRD and TV (Theorems 4 and 5), the empirical UOT plan (Theorem 12), and empirical barycenter (Theorems 16 and 17) but for the multinomial model in (6). We do not explicitly state the convergence statements for the empirical KRD (Theorem 7 and Theorem 8) as they can be immediately derived from our bound in (10). The proofs for all subsequent results only differ in the way the respective expectations are bounded, so whenever suitable, we only state relevant differences in the proofs. Notably, in Appendix E.3 we detail some simulations for several classes of measures which showcase that our theoretical results are realized.

Theorem 21 (Expected deviation of estimator in TV and KRD) *Let $(\mathcal{X}, d)$ be a finite metric space and $\mu \in \mathcal{M}_+(\mathcal{X})$ with total mass $\mathbb{M}(\mu)$. Let $\widehat{\mu}_N$ be the estimator from (6). Then, for any $p \geq 1$ it holds that*

$$\mathbb{E}\left[\mathrm{KR}^p_{p,C}\left(\widehat{\mu}_N, \mu\right)\right] \leq C^p \mathbb{E}\left[\mathrm{TV}\left(\widehat{\mu}_N, \mu\right)\right] \leq \left(C^p \sqrt{\mathbb{M}(\mu)} \sum_{x \in \mathcal{X}} \sqrt{\mu(x)}\right) N^{-\frac{1}{2}}.$$

Proof Following the proof of Theorem 4, it suffices to bound the TV norm in expectation,

$$\begin{aligned}\mathbb{E}\left[\mathrm{TV}\left(\widehat{\mu}_N, \mu\right)\right] &= \sum_{x \in \mathcal{X}} \mathbb{E}\left[\left|\, \widehat{\mu}_N(x) - \mu(x) \,\right|\right] = \frac{\mathbb{M}(\mu)}{N} \sum_{x \in \mathcal{X}} \mathbb{E}\Big[\Big| \sum_{i=1}^N \mathbb{1}\{X_i = x\} - N \frac{\mu(x)}{\mathbb{M}(\mu)}\Big|\Big] \\ &\leq \frac{\mathbb{M}(\mu)}{N} \sum_{x \in \mathcal{X}} \sqrt{N \frac{\mu(x)}{\mathbb{M}(\mu)} \left(1 - \frac{\mu(x)}{\mathbb{M}(\mu)}\right)} \leq N^{-\frac{1}{2}} \sqrt{\mathbb{M}(\mu)} \sum_{x \in \mathcal{X}} \sqrt{\mu(x)},\end{aligned}$$

where the inequality follows from the fact the $X_i \sim Ber\left(\mu(x)/\mathbb{M}(\mu)\right)$ for $i = 1, \ldots, N$. ■

Theorem 22 (Expected deviation of estimator in KRD) *Let $(\mathcal{X}, d)$ be a finite metric space and $\mu \in \mathcal{M}_+(\mathcal{X})$ with total mass $\mathbb{M}(\mu)$. Let $\widehat{\mu}_N$ be the estimator from (6). Then, for any $p \geq 1$, resolution $q > 1$ and depth $L \in \mathbb{N}$ it holds that*

$$\mathbb{E}\left[\mathrm{KR}_{p,C}\left(\widehat{\mu}_N, \mu\right)\right] \leq \mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C)^{1/p} N^{-\frac{1}{2p}}.$$

For

$$A_{q,p,L,\mathcal{X}}(l) := \mathrm{diam}(\mathcal{X})^p 2^{p-1} \left(q^{-Lp} |\mathcal{X}|^{\frac{1}{2}} + \left(\frac{q}{q-1}\right)^p \sum_{j=l}^{L} q^{p-jp} |Q_j|^{\frac{1}{2}}\right),$$

the constant is equal to

$$\mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C, q, L) = \begin{cases} \mathbb{M}(\mu) A_{q,p,L,\mathcal{X}}(1), & \text{if } C \geq 2h_{q,L}(0), \\ \mathbb{M}(\mu) A_{q,p,L,\mathcal{X}}(l), & \text{if } 2h_{q,L}(l) \leq C < 2h_{q,L}(l-1), \\ \frac{C^p}{2} \sqrt{\mathbb{M}(\mu)} \sum_{x \in \mathcal{X}} \sqrt{\mu(x)}, & \text{if } C \leq \left(2h_{q,L}(L) \vee \min_{x \neq x'} d(x, x')\right). \end{cases}$$

Furthermore, for $p = 1$ the factor $\frac{q}{(q-1)}$ in $A_{q,1,L,\mathcal{X}}(a, b, l)$ can be removed. Denote

$$\mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C) := \inf_{L\in\mathbb{N},q>1} \mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C, q, L).$$

Proof The proof of this result only differs from the proof of Theorem 5 by the upper bounds on the relevant expectations. By definition $\mathbb{E}\left[|\mathbb{M}(\widehat{\mu}_N) - \mathbb{M}(\mu)|\right] = 0$. Furthermore, scaling the expectation by total mass

$$\mathbb{E}\left[\left|\widehat{\mu}^L_N(\mathcal{C}(x)) - \mu^L(\mathcal{C}(x))\right|\right] = \mathbb{M}(\mu)\mathbb{E}\left[\left|\frac{\widehat{\mu}^L_N(\mathcal{C}(x))}{\mathbb{M}(\mu)} - \frac{\mu^L(\mathcal{C}(x))}{\mathbb{M}(\mu)}\right|\right],$$

we notice that $\frac{\widehat{\mu}^L_N(\mathcal{C}(x))}{\mathbb{M}(\mu)} \overset{D}{=} \frac{1}{N}\sum_{i=1}^N X_i(x)$, where $X_1(x), \dots, X_N(x) \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}\,(a(x))$ with $a(x) \coloneqq \frac{\mu^L(\mathcal{C}(x))}{\mathbb{M}(\mu)}$. Consequently, it holds that

$$\begin{aligned}
\sum_{x\in Q_l} \mathbb{E}\left[\left|\widehat{\mu}^L_N(\mathcal{C}(x)) - \mu^L(\mathcal{C}(x))\right|\right] &= \mathbb{M}(\mu) \sum_{x\in Q_l} \mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N X_i(x) - a(x)\right|\right] \\
&\le \mathbb{M}(\mu) \sum_{x\in Q_l} \sqrt{\frac{a(x)(1-a(x))}{N}} \\
&\le \mathbb{M}(\mu)\sqrt{\frac{|Q_l|}{N}}.
\end{aligned}$$

■

Notably, compared to $\mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C, q, L)$, the constant $\mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C, q, L)$ misses an additional summand for large C. This summand corresponds to the estimation error of the total mass intensity of $\widehat{\mu}_N$ which is zero by assumptions of the model.

Remark 23 *If $C > \mathrm{diam}(\mathcal{X})$ and $\mathbb{M}(\mu) = \mathbb{M}(\nu)$, the (p, C)-KRD between μ and ν is equal to the Wasserstein distance between these two measures. In particular, for $C > 2h_{q,L}(0)$ we recover the respective deviation bounds by Sommerfeld et al. (2019) for the measure estimator Wasserstein distance. In particular, for this setting, it holds in the multinomial model for all $N \in \mathbb{N}$ that $\mathbb{M}(\widehat{\mu}_N) = \mathbb{M}(\mu)$ which implies for $C > \mathrm{diam}(\mathcal{X})$ that $\mathrm{KR}_{p,C}(\widehat{\mu}_N, \mu) = W_p(\widehat{\mu}_N, \mu)$. Since for the latter term the parametric rate $N^{-\frac{1}{2p}}$ is already known to be optimal (Sommerfeld et al., 2019), our rate in N is sharp.*

Theorem 24 (Expected deviation of empirical UOT plans) *Let $(\mathcal{X}, d)$ be a finite metric space and $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$. Let $\widehat{\mu}_N, \widehat{\nu}_N$ be their respective estimators from (6). Then, for any $p \ge 1$ and $C \ge 0$ it follows that*

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{H}_{\mathrm{TV}}\left(\mathbf{P}^*_{p,C}(\widehat{\mu}_N, \widehat{\nu}_N), \mathbf{P}^*_{p,C}(\mu, \nu)\right)\right] \\
&\le 4(|\mathcal{X}| + 1)N^{-1/2}\left(\sqrt{\mathbb{M}(\mu)}\sum_{x\in\mathcal{X}}\sqrt{\mu(x)} + \sqrt{\mathbb{M}(\nu)}\sum_{x\in\mathcal{X}}\sqrt{\nu(x)}\right) \\
&\le 4(|\mathcal{X}| + 1)^{3/2}N^{-1/2}\mathbb{M}(\mu + \nu).
\end{aligned}$$

Proof The assertion follows by combining our stability bound (Theorem 11) with convergence of the estimators $\widehat{\mu}_N, \widehat{\nu}_N$ to μ, ν with respect to the total variation norm (Theorem 21) and the fact that by definition $\mathbb{E}\left[|\mathbb{M}(\widehat{\mu}_N) - \mathbb{M}(\mu)|\right] = \mathbb{E}\left[|\mathbb{M}(\widehat{\nu}_N) - \mathbb{M}(\nu)|\right] = 0$. The last inequality is due to Cauchy–Schwarz. ■

Remark 25 *The convergence rate in N matches with the distributional limit obtained by Klatt et al. (2022) and Liu et al. (2023) and is therefore sharp in N. In particular, our result explicitly quantifies how the number of support points affect the convergence rate.*

Theorem 26 (Expected deviation of Fréchet error) *Let $\mu^1, \ldots, \mu^J \in \mathcal{M}_+(\mathcal{X})$ and denote $\mathcal{X}_i = \operatorname{supp}(\mu^i)$ for $i = 1, \ldots, J$. Consider random estimators $\widehat{\mu}^1_{N_i}, \ldots, \widehat{\mu}^J_{N_J} \in \mathcal{M}_+(\mathcal{X})$ derived from (6) and based on sample size $N_1, \ldots, N_J$, respectively. Then it holds for any barycenter μ^* of the population measures and any barycenter $\widehat{\mu}^*$ of the estimators,*

$$\mathbb{E}\left[|F_{p,C}(\widehat{\mu}^\star) - F_{p,C}(\mu^\star)|\right] \leq \frac{2pC^{p-1}}{J} \sum_{i=1}^{J} \mathcal{E}^{\mathrm{Mult}}_{1,\mathcal{X}_i,\mu^i}(C) N_i^{-\frac{1}{2}}.$$

Theorem 27 (Expected deviation of empirical barycenters) *Let $\mu^1, \ldots, \mu^J \in \mathcal{M}_+(\mathcal{X})$ and denote $\mathcal{X}_i = \operatorname{supp}(\mu^i)$ for $i = 1, \ldots, J$. Consider random estimators $\widehat{\mu}^1_{N_i}, \ldots, \widehat{\mu}^J_{N_J} \in \mathcal{M}_+(\mathcal{X})$ derived from (6) and based on sample size $N_1, \ldots, N_J$, respectively. Let $\mathbf{B}^\star$ be the set of (p, C)-barycenters of $\mu^1, \ldots, \mu^J$ and $\widehat{\mathbf{B}}^\star$ the set of (p, C)-barycenters of $\widehat{\mu}^1_{N_i}, \ldots, \widehat{\mu}^J_{N_J}$. Then, for $p \geq 1$ it holds that*

$$\mathbb{E}\left[\sup_{\widehat{\mu}^\star \in \widehat{\mathbf{B}}^\star} \inf_{\mu^\star \in \mathbf{B}^\star} \mathrm{KR}^p_{p,C}(\widehat{\mu}^\star, \mu^\star)\right] \leq \frac{pC^{p-1}}{V_P J} \sum_{i=1}^{J} \mathcal{E}^{\mathrm{Mult}}_{1,\mathcal{X}_i,\mu^i}(C) N_i^{-\frac{1}{2}},$$

where the constant V_P is defined as in Theorem 17.

The proofs of Theorem 26 and Theorem 27 are deferred to Appendix D.5.

Remark 28 *By the same argument as for the Poisson model (Theorem 19), the convergence rate for the empirical barycenter does not to zero faster than for a single measure. Thus, the $N^{-1/2}$ rate is sharp.*

In the following we, derive explicit bounds for $\mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C)$ for under the structural Assumption (19). To this end, we follow the arguments in Section 2.4. For the regime, for $\alpha < 2p$ and $L \to \infty$ we obtain,

$$\mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C) \leq \begin{cases} \mathbb{M}(\mu) A^{1/2} \mathrm{diam}(\mathcal{X})^p 2^{3p-1} \frac{2^{\alpha/2-p}}{1-2^{\alpha/2-p}}, \\ \qquad \text{if } C \geq 2h_L(0), \\ \mathbb{M}(\mu) A^{1/2} \mathrm{diam}(\mathcal{X})^p 2^{3p-1} \frac{2^{(\alpha/2-p)l}}{1-2^{\alpha/2-p}}, \\ \qquad \text{if } 2h_L(l) \leq C < 2h_L(l-1), \\ \frac{C^p}{2} \sqrt{\mathbb{M}(\mu)} \sum_{x \in \mathcal{X}} \sqrt{\mu(x)}, \\ \qquad \text{if } C \leq \left(2h_L(L) \vee \min_{x \neq x'} d(x, x')\right). \end{cases}$$

For $\alpha = 2p$ and $L = \lfloor \frac{1}{\alpha} \log_2(|\mathcal{X}|) \rfloor$ we get

$$\mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C) \leq \begin{cases} \mathbb{M}(\mu)\mathrm{diam}(\mathcal{X})^p 2^{3p-1} \left(2^{-p} + A^{1/2} \frac{1}{\alpha} \log_2 |\mathcal{X}|\right), \\ \qquad \text{if } C \geq 2h_L(0), \\ \mathbb{M}(\mu)\mathrm{diam}(\mathcal{X})^p 2^{3p-1} (2^{-p} + A^{1/2} \left(\frac{1}{\alpha} \log_2 |\mathcal{X}| - l\right)), \\ \qquad \text{if } 2h_L(l) \leq C < 2h_L(l-1), \\ \frac{C^p}{2} \sqrt{\mathbb{M}(\mu)} \sum_{x \in \mathcal{X}} \sqrt{\mu(x)}, \\ \qquad \text{if } C \leq \left(2h_L(L) \vee \min_{x \neq x'} d(x, x')\right). \end{cases}$$

Finally, for $\alpha > 2p$ and $L = \lfloor \frac{1}{\alpha} \log_2(|\mathcal{X}|) \rfloor$, it holds

$$\mathcal{E}^{\mathrm{Mult}}_{p,\mathcal{X},\mu}(C) \leq \begin{cases} \mathbb{M}(\mu)\mathrm{diam}(\mathcal{X})^p 2^{3p-1} |\mathcal{X}|^{1/2-p/\alpha} \left(2^{-p} + A^{1/2} \frac{2^{\alpha/2-2p}}{2^{\alpha/2-p}-1}\right), \\ \qquad \text{if } C \geq 2h_L(0), \\ \mathbb{M}(\mu)\mathrm{diam}(\mathcal{X})^p 2^{3p-1} \times \\ \left(2^{-p} |\mathcal{X}|^{1/2-p/\alpha} + A^{1/2} \frac{2^{\alpha/2-p}}{2^{\alpha/2-p}-1} \left(|\mathcal{X}|^{1/2-p/\alpha} - 2^{(\alpha/2-p)(l-1)}\right)\right), \\ \qquad \text{if } 2h_L(l) \leq C < 2h_L(l-1), \\ \frac{C^p}{2} \sqrt{\mathbb{M}(\mu)} \sum_{x \in \mathcal{X}} \sqrt{\mu(x)}, \\ \qquad \text{if } C \leq \left(2h_L(L) \vee \min_{x \neq x'} d(x, x')\right). \end{cases}$$

We stress that while these constants do not include the additional term for the estimation of the total mass intensity, their dependency on $|\mathcal{X}|$ is identical to that of the upper bounds on $\mathcal{E}^{\mathrm{Pois}}_{p,\mathcal{X},\mu}(C)$. In particular, the phase transitions still occur depending on whether α is larger than $2p$, smaller than $2p$ or equal to it.

Appendix B. Bounds for the Bernoulli Model

In this section we provide results analogue to previous subsection for the estimator in the Bernoulli model in (7). As before, since the proofs only differ in the way expectations are bounded, we only state relevant differences in the proofs. Simulations, corroborating our theoretical findings are provided in Appendix E.4.

Theorem 29 (Expected deviation of estimator in TV and KRD) *Let $(\mathcal{X}, d)$ be a finite metric space and $\mu \in \mathcal{M}_+(\mathcal{X})$ with $\mu(x) \in \{0, 1\}$ for $x \in \mathcal{X}$. Let $\widehat{\mu}_{s_\mathcal{X}}$ be the measure in (7). Then, for any $p \geq 1$ it holds that*

$$\mathbb{E}\left[\mathrm{KR}^p_{p,C}\left(\widehat{\mu}_{s_\mathcal{X}}, \mu\right)\right] \leq C^p \mathbb{E}\left[\mathrm{TV}\left(\widehat{\mu}_{s_\mathcal{X}}, \mu\right)\right] \leq 2C^p \sum_{x \in \mathcal{X}} (1 - s_x).$$

Proof This proof is identical to the proof of Theorem 4 except for the bound on the expectation. For this, note that

$$\mathbb{E}\left[\mathrm{TV}\left(\widehat{\mu}_{s_\mathcal{X}}, \mu\right)\right] = \sum_{x \in \mathcal{X}} \mathbb{E}\left[\left|\frac{1}{s_x} B_x - \mu(x)\right|\right] \leq \sum_{x \in \mathrm{supp}(\mu)} (1 - s_x) + s_x\left(\frac{1}{s_x} - 1\right) \leq 2 \sum_{x \in \mathcal{X}} (1 - s_x),$$

with $B_x \sim \mathrm{Ber}(s_x \mu(x))$ for $s_x \in (0,1]$ for all $x \in \mathcal{X}$. ■

Theorem 30 (Expected deviation of estimator in KRD) *Let $(\mathcal{X}, d)$ be a finite metric space and $\mu \in \mathcal{M}_+(\mathcal{X})$ with $\mu(x) \in \{0,1\}$ for $x \in \mathcal{X}$. Let $\widehat{\mu}_{s_\mathcal{X}}$ be the measure in (7). Then, for any $p \geq 1$, resolution $q > 1$ and depth $L \in \mathbb{N}$ it holds that*

$$\mathbb{E}\left[\mathrm{KR}_{p,C}\left(\widehat{\mu}_{s_\mathcal{X}}, \mu\right)\right] \leq \mathcal{E}^{\mathrm{Ber}}_{p,\mathcal{X},\mu}(C,q,L)^{1/p} \begin{cases} \left(2\sum_{x\in\mathcal{X}}(1-s_x)\right)^{\frac{1}{p}}, \\ \qquad\qquad \text{if } C \leq \left(2h_{q,L}(L) \vee \min_{x\neq x'} d(x,x')\right) \\ \left(\sum_{x\in\mathcal{X}} \frac{1-s_x}{s_x}\right)^{\frac{1}{2p}}, \\ \qquad\qquad \text{else.} \end{cases}$$

The constant $\mathcal{E}^{\mathrm{Ber}}_{p,\mathcal{X},\mu}(C,q,L)$ is equal to $\mathcal{E}^{\mathrm{Poi}}_{p,\mathcal{X},\mu}(C,q,L)$ for all $C > 0$, $q > 1$ and $L \in \mathbb{N}$. We denote

$$\mathcal{E}^{\mathrm{Ber}}_{p,\mathcal{X},\mu}(C) := \inf_{L\in\mathbb{N}, q>1} \mathcal{E}^{\mathrm{Ber}}_{p,\mathcal{X},\mu}(C,q,L).$$

Proof The proof of this result only differs from the proof of Theorem 5 in the upper bounds on the relevant expectations. Recall the estimator $\widehat{\mu}_{s_\mathcal{X}}$ from (7) and let $B_x \sim \mathrm{Ber}(s_x\mu(x))$ for $s_x \in (0,1]$ for all $x \in \mathcal{X}$. It holds that

$$\begin{aligned} \sum_{x\in Q_l} \mathbb{E}\left[\left|\widehat{\mu}^L_{s_\mathcal{X}}(\mathcal{C}(x)) - \mu^L(\mathcal{C}(x))\right|\right] &= \sum_{x\in Q_l} \mathbb{E}\left[\left|\sum_{y\in\mathcal{C}(x)} \frac{B_y}{s_y} - \sum_{y\in\mathcal{C}(x)} \mu^L(y)\right|\right] \\ &\leq \sum_{x\in Q_l} \sqrt{\mathrm{Var}\left(\sum_{y\in\mathcal{C}(x)} \frac{B_y}{s_y}\right)} = \sum_{x\in Q_l} \sqrt{\sum_{y\in\mathcal{C}(x)} s_y^{-2}\mathrm{Var}(B_y)} \\ &= \sum_{x\in Q_l} \sqrt{\sum_{y\in\mathcal{C}(x)} \frac{(1-\mu(y)s_y)\mu(y)}{s_y}} \leq \sqrt{|Q_l|}\sqrt{\sum_{x\in\mathcal{X}} \frac{1-s_x}{s_x}}. \end{aligned}$$

The total mass can be bounded analogously as

$$\mathbb{E}\left[\left|\widehat{\mu}^L_{s_\mathcal{X}}(\mathcal{X}) - \mu^L(\mathcal{X})\right|\right] \leq \sqrt{\sum_{x\in\mathcal{X}} \frac{1-s_x}{s_x}}. \tag{26}$$

■

Since the constants for the deviation bounds for this model coincide with those for the Poisson model we refer to the previous discussion on their properties.

Remark 31 *Consider $s_\mathcal{X}$ such that $s_x = s$ for some $s \in (0,1]$ and all $x \in \mathcal{X}$. Note, that for sufficiently small C the upper bound is an equality, since the (p,C)-KRD in this setting is proportional to the TV distance and that distance has a closed form solution here. For*

larger C, the expectation in the proof of Theorem 30 amounts to bounding the mean absolute deviation of a binomial distribution. This has a closed form solution which scales as the standard deviation of the respective binomial for s not too close to 0 or 1 (Berend and Kontorovich, 2013). Hence, in this context the upper bound on the mean absolute deviation in the proof is sharp. So based on the presented approach for the deviation bounds, the upper bound is non-improvable.

To state the result analogous to Theorem 11, we consider μ and $\nu \in \mathcal{M}_+(\mathcal{X})$ with respective supports $\mathcal{X}_\mu := \operatorname{supp}(\mu) = \{x_1, \dots, x_M\}$ and success probabilities $s_x \in [0,1]$ for $x \in \mathcal{X}_\mu$, and $\mathcal{X}_\nu := \operatorname{supp}(\nu) = \{x'_1, \dots, x'_{M'}\}$ with success probabilities $s_{x'} \in [0,1]$ for $x' \in \mathcal{X}_\nu$.

Theorem 32 (Expected deviation of empirical UOT plans) *Let $(\mathcal{X}, d)$ be a finite metric space and $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$. Let $\widehat{\mu}_{s_{\mathcal{X}_\mu}}, \widehat{\nu}_{s_{\mathcal{X}_\nu}}$ be their respective estimators from (7). Then, for any $p \geq 1$ and $C \geq 0$ it follows that*

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{H}_{\mathrm{TV}}\left(\mathbf{P}^*_{p,C}(\widehat{\mu}_{s_{\mathcal{X}_\mu}}, \widehat{\nu}_{s_{\mathcal{X}_\nu}}), \mathbf{P}^*_{p,C}(\mu,\nu)\right)\right] \\
&\leq 4(|\mathcal{X}|+1)\left(2\sum_{x\in\mathcal{X}_\mu}(1-s_x) + 2\sum_{x'\in\mathcal{X}_\nu}(1-s_{x'}) + \sqrt{\sum_{x\in\mathcal{X}_\mu}\frac{1-s_x}{s_x}} + \sqrt{\sum_{x'\in\mathcal{X}_\nu}\frac{1-s_{x'}}{s_{x'}}}\right) \\
&\leq 4(|\mathcal{X}|+1)\left(2(\mathbb{M}(\mu+\nu)-2) + \sqrt{\sum_{x\in\mathcal{X}_\mu}\frac{1-s_x}{s_x}} + \sqrt{\sum_{x'\in\mathcal{X}_\nu}\frac{1-s_{x'}}{s_{x'}}}\right)
\end{aligned}$$

Proof The assertion follows by combining our stability bound (Theorem 11) with convergence of the estimators $\widehat{\mu}_{s_{\mathcal{X}_\mu}}, \widehat{\nu}_{s_{\mathcal{X}_\nu}}$ to μ, ν with respect to the total variation norm and in terms of their masses (26). ■

Theorem 33 (Expected deviation of Fréchet error) *Let $\mu^1, \dots, \mu^J \in \mathcal{M}_+(\mathcal{X})$ and denote $\mathcal{X}_i = \operatorname{supp}(\mu^i)$ for $i = 1, \dots, J$. Consider (random) estimators $\widehat{\mu}^1_{s_{\mathcal{X}_1}}, \dots, \widehat{\mu}^J_{s_{\mathcal{X}_J}} \in \mathcal{M}_+(\mathcal{X})$ derived from (7). Then,*

$$\mathbb{E}\left[|F_{p,C}(\widehat{\mu}^\star) - F_{p,C}(\mu^\star)|\right] \leq \frac{2pC^{p-1}}{J}\sum_{i=1}^{J}\mathcal{E}^{\mathrm{Ber}}_{1,\mathcal{X}_i,\mu^i}(C)\psi(s_{\mathcal{X}_i}),$$

where ψ is given by

$$\psi(s_{\mathcal{X}}) = \begin{cases} (2\sum_{x\in\mathcal{X}}(1-s_x), & \text{if } C \leq \min_{x\neq x'} d(x,x') \\ \left(\sum_{x\in\mathcal{X}}\frac{1-s_x}{s_x}\right)^{\frac{1}{2}}, & \text{else.} \end{cases}$$

Theorem 34 (Expected deviation of empirical barycenters) *Let $\mu^1, \dots, \mu^J \in \mathcal{M}_+(\mathcal{X})$ and denote $\mathcal{X}_i = \operatorname{supp}(\mu^i)$ for $i = 1, \dots, J$. Consider (random) estimators $\widehat{\mu}^1_{s_{\mathcal{X}_1}}, \dots, \widehat{\mu}^J_{s_{\mathcal{X}_J}} \in$*

$\mathcal{M}_+(\mathcal{X})$ *derived from* (7). *Let* $\mathbf{B}^\star$ *be the set of* (p, C)*-barycenters of* $\mu^1, \dots, \mu^J$ *and* $\widehat{\mathbf{B}}^\star$ *the set of* (p, C)*-barycenters of* $\widehat{\mu}^1_{s_{\mathcal{X}_1}}, \dots, \widehat{\mu}^J_{s_{\mathcal{X}_J}}$. *Then, for* $p \geq 1$ *it holds that*

$$\mathbb{E}\left[\sup_{\widehat{\mu}^\star \in \widehat{\mathbf{B}}^\star} \inf_{\mu^\star \in \mathbf{B}^\star} \mathrm{KR}^p_{p,C}(\widehat{\mu}^\star, \mu^\star)\right] \leq \frac{pC^{p-1}}{V_P J} \sum_{i=1}^{J} \mathcal{E}^{\mathrm{Ber}}_{1,\mathcal{X}_i,\mu^i}(C)\psi(s_{\mathcal{X}_i}),$$

where ψ *is defined as in Theorem 33 and* V_P *is defined as in Theorem 16.*

The proofs of Theorem 33 and Theorem 34 are deferred to Appendix D.5.

Appendix C. A Lift to the Balanced Optimal Transport Problem

A key tool in establishing properties of the (p, C)-KRD and the (p, C)-barycenter is the lift of these problems to the space of probability measures by augmenting the space $\mathcal{X}$ with a dummy point having a fixed distance to all points in $\mathcal{X}$, see (Heinemann et al., 2023, Section 3.1) for details. For a fixed parameter $C > 0$, consider a dummy point $\mathfrak{d}$ and define the augmented space $\widetilde{\mathcal{X}} \coloneqq \mathcal{X} \cup \{\mathfrak{d}\}$ with metric cost

$$\widetilde{d}^p_C(x, x') = \begin{cases} d^p(x, x') \wedge C^p, & \text{if } x, x' \in \mathcal{X}, \\ \frac{C^p}{2}, & \text{if } x \in \mathcal{X},\ x' = \mathfrak{d}, \\ \frac{C^p}{2}, & \text{if } x = \mathfrak{d},\ x' \in \mathcal{X}, \\ 0, & \text{if } x = x' = \mathfrak{d}. \end{cases} \tag{27}$$

Consider the subset $\mathcal{M}^B_+(\mathcal{X}) \coloneqq \{\mu \in \mathcal{M}_+(\mathcal{X}) \mid \mathbb{M}(\mu) \leq B\} \subset \mathcal{M}_+(\mathcal{X})$ of non-negative measures whose total mass is bounded by B. Setting $\widetilde{\mu} \coloneqq \mu + (B - \mathbb{M}(\mu))\delta_{\mathfrak{d}}$, any measure $\mu \in \mathcal{M}^B_+(\mathcal{X})$ defines an *augmented measure* $\widetilde{\mu}$ on $\mathcal{X}$ such that $\mathbb{M}(\widetilde{\mu}) = B$. For any $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ and their augmented versions $\widetilde{\mu}, \widetilde{\nu} \in \mathcal{M}_+(\mathcal{X})$ it holds

$$\mathrm{KR}^p_{C,p}(\mu, \nu) = \widetilde{\mathrm{OT}}_{\widetilde{d}^p_C}(\widetilde{\mu}, \widetilde{\nu}). \tag{28}$$

Here, $\widetilde{\mathrm{OT}}_{\widetilde{d}^p_C}$ denotes the OT cost defined for measures μ, ν on $(\widetilde{\mathcal{X}}, \widetilde{d})$ with $\mathbb{M}(\mu) = \mathbb{M}(\nu)$ as

$$\widetilde{\mathrm{OT}}_{\widetilde{d}^p_C}(\mu, \nu) \coloneqq \min_{\pi \in \Pi_=(\mu,\nu)} \sum_{x,x' \in \mathcal{X}} \widetilde{d}^p_C(x, x')\pi(x, x'),$$

where the set of couplings $\Pi_=(\mu, \nu)$ is the set $\Pi_\leq(\mu, \nu)$ with inequalities replaced by equalities.

Similarly, the (p, C)-barycenter problem can be augmented. For this, let $\widetilde{\mathcal{Y}} \coloneqq \mathcal{Y} \cup \{\mathfrak{d}\}$ endowed with the metric $\widetilde{d}_C$ in (27) (replace $\mathcal{X}$ by $\mathcal{Y}$ and recall that $\mathcal{X} \subset \mathcal{Y}$) and augment the measures $\mu^1, \dots, \mu^J$ to $\widetilde{\mu}^1, \dots, \widetilde{\mu}^J$ where $\widetilde{\mu}_i = \mu^i + \sum_{j \neq i} \mathbb{M}(\mu^i)\delta_{\mathfrak{d}}$ for $1 \leq i \leq J$. In particular, it holds $\mathbb{M}(\widetilde{\mu}_i) = \sum_{i=1}^J \mathbb{M}(\mu^i)$ and the *augmented p-Fréchet functional* is defined as

$$\widetilde{F}_{p,C}(\mu) \coloneqq \frac{1}{J} \sum_{i=1}^{J} \widetilde{\mathrm{OT}}_{\widetilde{d}^p_C}(\widetilde{\mu}_i, \mu).$$

Any minimizer of $\widetilde{F}_{p,C}$ is referred to as augmented (p, C)-barycenter.

LP-Formulation for the $(\mathbf{p}, \mathbf{C})$-Barycenter

According to (Heinemann et al., 2023, Lemma 3.2), the augmented (p, C)-barycenter problem can be rewritten as a linear program based on the centroid set $\widetilde{\mathcal{C}}_{KR}(J, p, C) = \mathcal{C}_{KR}(J, p, C) \cup \{\mathfrak{d}\}$ (recall (5) for the definition of $\mathcal{C}_{KR}(J, p, C)$) of the augmented measures. This yields

$$
\begin{aligned}
\min_{\pi^{(1)},\dots,\pi^{(J)},a} \quad & \frac{1}{J}\sum_{i=1}^{J} \sum_{j=1}^{|\widetilde{\mathcal{C}}_{\mathrm{KR}}(J,p,C)|} \sum_{k=1}^{M_i} \pi_{jk}^{(i)} c_{jk}^{i} \\
\text{s.t.} \quad & \sum_{k=1}^{M_i} \pi_{jk}^{(i)} = a_j, && \forall\ i = 1,\dots,J, \forall j = 1,\dots,|\widetilde{\mathcal{C}}_{\mathrm{KR}}(J,p,C)|, \\
& \sum_{j=1}^{|\widetilde{\mathcal{C}}_{\mathrm{KR}}(J,p,C)|} \pi_{jk}^{(i)} = b_k^i, && \forall\ i = 1,\dots,J, \forall k = 1,\dots,M_i, \\
& \pi_{jk}^{(i)} \geq 0, && \forall i = 1,\dots,J, \forall j = 1,\dots,|\widetilde{\mathcal{C}}_{\mathrm{KR}}(J,p,C)|, \\
& && \forall k = 1,\dots,M_i,
\end{aligned}
\tag{29}
$$

where $M_i = |\widetilde{\mathcal{X}}_i|$ for each $1 \leq i \leq J$ is the cardinality of the support of the augmented measure $\widetilde{\mu}_i$. Here, c_{jk}^i denotes the distance between the j-th point of $|\widetilde{\mathcal{C}}_{\mathrm{KR}}(J, p, C)|$ and the k-th point in the support of $\widetilde{\mu}_i$, while b^i is the vector of masses corresponding to $\widetilde{\mu}_i$.

Appendix D. Auxiliary Results and Deferred Proofs

D.1 Proof of Stability Bound of KRD via Dual Formulation

Proof [Proof of Lemma 3] We first observe that the second assertion follows from $|x - y| \leq x^{1-p}\,|x^p - y^p|$ for all $x, y \geq 0, p \geq 1$. To prove (17), note that if $\mu^i = 0$ or $\nu^i = 0$ for each $i \in \{1, 2\}$, then by $\Pi_{\leq}(\mu^i, \nu^i) = \{0\}$ it follows that

$$
\mathrm{KR}_{p,C}(\mu^i, \nu^i) = \frac{C^p}{2}\left(\mathbb{M}(\mu^i) + \mathbb{M}(\nu^i)\right), \tag{30}
$$

and the assertion follows by triangle inequality,

$$
\begin{aligned}
|\mathrm{KR}_{p,C}^p(\mu^1, \nu^1) - \mathrm{KR}_{p,C}^p(\mu^2, \nu^2)| &= \frac{C^p}{2}\Big|\mathbb{M}(\mu^1) + \mathbb{M}(\nu^1) - \mathbb{M}(\mu^2) - \mathbb{M}(\nu^2)\Big| \\
&\leq \frac{C^p}{2}\left(\big|\mathbb{M}(\mu^1) - \mathbb{M}(\mu^2)\big| + \big|\mathbb{M}(\nu^1) - \mathbb{M}(\nu^2)\big|\right).
\end{aligned}
$$

Moreover, if $C \leq \min\{\Delta(\mu^1, \nu^1), \Delta(\mu^2, \nu^2)\}$, then by Heinemann et al. (2023, Lemma 2.1) the UOT plan between μ^i and ν^i is to not transport anything, and delete the excess mass at cost $C^p/2$, which yields (30) and again inequality (17) follows by triangle inequality.

For the remaining case, we augment μ^i, ν^i to $\widetilde{\mu}^i$, $\widetilde{\nu}^i$ on $\widetilde{\mathcal{X}} = \mathcal{X} \cup \{\mathfrak{d}\}$ as spelled out in Appendix C with total mass $B \coloneqq \max(\mathbb{M}(\mu^i), \mathbb{M}(\nu^i))$. Then, by the representation of the

UOT cost in terms of the augmented OT cost (see Equation (28)), it follows that

$$\mathrm{KR}^p_{p,C}(\mu^i, \nu^i) = \widetilde{\mathrm{OT}}_{\widetilde{d}^p_C}(\widetilde{\mu}^i, \widetilde{\nu}^i) = \max_{\substack{f,g:\widetilde{\mathcal{X}}\to\mathbb{R} \\ f(x)+g(x')\leq \widetilde{d}^p_C(x,x')}} \sum_{x\in\widetilde{\mathcal{X}}} f(x)\widetilde{\mu}^i(x) + \sum_{x'\in\widetilde{\mathcal{X}}} g(x')\widetilde{\nu}^i(x').$$

By Villani (2003, Remark 1.13) there always exist optimal potentials f and g such that they are bounded by $\|\widetilde{d}^p_C\|_\infty \leq C^p$ in absolute value. Hence, upon denoting such pairs of optimal potentials for $\widetilde{\mathrm{OT}}^p_{p,C}(\mu^i, \nu^i)$ by (f_i, g_i), it follows that

$$\begin{aligned}
\mathrm{KR}^p_{p,C}(\mu^1, \nu^1) - \mathrm{KR}^p_{p,C}(\mu^2, \nu^2) &\leq \sum_{x\in\mathcal{X}} f_1(x)(\widetilde{\mu}^1 - \widetilde{\mu}^2)(x) + \sum_{x'\in\widetilde{\mathcal{X}}} g_1(x')(\widetilde{\nu}^1 - \widetilde{\nu}^2)(x') \\
&\leq C^p \left(\sum_{x\in\widetilde{\mathcal{X}}} |(\widetilde{\mu}^1 - \widetilde{\mu}^2)(x)| + \sum_{x\in\widetilde{\mathcal{X}}} |(\widetilde{\mu}^1 - \widetilde{\mu}^2)(x)| \right) \\
&= C^p \left(\mathrm{TV}(\widetilde{\mu}^1, \widetilde{\mu}^2) + \mathrm{TV}(\widetilde{\nu}^1, \widetilde{\nu}^2) \right) \\
&\leq 2C^p \left(\mathrm{TV}(\mu^1, \mu^2) + \mathrm{TV}(\nu^1, \nu^2) \right),
\end{aligned}$$

where the last inequality follows from Lemma 38. Exchanging (μ^1, ν^1) with (μ^2, ν^2) yields a corresponding lower bound and finishes the proof. ■

D.2 Proof of Statistical Deviation Bound for Empirical KRD via TV-norm and Tree Approximation

As a preparatory step to prove Theorem 5, we treat the significantly simpler case of an empirical deviation bound with respect to the total variation distance.

Proof [Proof of Theorem 4] Let $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$. By retaining all common mass between μ and ν at place and delete (resp. create) excess mass (resp. deficient mass) we obtain a feasible solution for (2) with objective value in terms of a total variation distance between μ and ν. Thus, it holds

$$\mathrm{KR}^p_{p,C}(\mu, \nu) \leq C^p \mathrm{TV}(\mu, \nu).$$

In particular, this inequality is satisfied for $\nu = \widehat{\mu}_{t,s}$. Taking expectations yields

$$\begin{aligned}
\mathbb{E}\left[\mathrm{TV}(\widehat{\mu}_{t,s}, \mu)\right] &= \frac{1}{st} \sum_{x\in\mathcal{X}} \mathbb{E}\left[|P_x B_x - st\mu(x)|\right] \\
&= \frac{1}{st} \sum_{x\in\mathcal{X}} s\mathbb{E}\left[|P_x - st\mu(x)|\right] + (1-s)st\mu(x) \\
&\leq \frac{1}{st} \sum_{x\in\mathcal{X}} s(1-s)\mathbb{E}\left[P_x\right] + s^2\mathbb{E}\left[|P_x - t\mu(x)|\right] + (1-s)st\mu(x) \\
&\leq \frac{1}{st} \sum_{x\in\mathcal{X}} 2s(1-s)t\mu_x + s^2\sqrt{t}\sqrt{\mu(x)} \\
&= 2(1-s)\mathbb{M}(\mu) + \frac{s}{\sqrt{t}} \sum_{x\in\mathcal{X}} \sqrt{\mu(x)}.
\end{aligned}$$

■

With Lemma 1 and Theorem 4 at our disposal we are able to prove Theorem 5.

Proof [Proof of Theorem 5] Let $\widehat{\mu}_{t,s}$ be the estimator from (8). We fix $p = 1$ and detail the case $p > 1$ at the end of the proof. Suppose first that $C \leq \min_{x \neq x'} d(x, x')$. According to (Heinemann et al., 2023, Theorem 2.2 (ii)) it holds that

$$\mathbb{E}\left[\mathrm{KR}_{1,C}(\widehat{\mu}_{t,s}, \mu)\right] = \frac{C}{2}\mathbb{E}\left[\sum_{x \in \mathcal{X}} |\widehat{\mu}_{t,s} - \mu(x)|\right] = \frac{C}{2}\mathbb{E}\left[\mathrm{TV}(\widehat{\mu}_{t,s}, \mu)\right].$$

This yields the total variation bounds (see Theorem 4). Next, consider the tree approximation as outlined in Section 2.1.1 and construct an ultrametric tree $\mathcal{T}$ such that $\mathrm{KR}_{1,C}(\widehat{\mu}_{t,s}, \mu) \leq \mathrm{KR}_{d_{\mathcal{T}},C}\left(\widehat{\mu}^L_{t,s}, \mu^L\right)$. Applying Lemma 1 for $p = 1$ where by definition the difference of height function is equal to

$$h_{q,L}(j-1) - h_{q,L}(j) = \frac{\mathrm{diam}(\mathcal{X})}{q-1}\left(q^{2-j} - q^{1-j}\right) = \mathrm{diam}(\mathcal{X}) q^{1-j}$$

and yields the upper bound

$$\mathbb{E}\left[\mathrm{KR}_{1,C}(\widehat{\mu}_{t,s}, \mu)\right] \leq \mathbb{E}\left[\mathrm{KR}_{d_{\mathcal{T}},C}\left(\widehat{\mu}^L_{t,s}, \mu^L\right)\right]$$

$$= \begin{cases} \left(\frac{C}{2} - h_{q,L}(0)\right)\mathbb{E}\left[|\mathbb{M}(\widehat{\mu}_{t,s}) - \mathbb{M}(\mu)|\right] & \\ +\mathrm{diam}(\mathcal{X}) \sum_{j=1}^{L+1} q^{1-j} \sum_{x \in Q_j} \mathbb{E}\left[\left|\widehat{\mu}^L_{t,s}(\mathcal{C}(x)) - \mu^L(\mathcal{C}(x))\right|\right], & \text{if } C \geq 2h_{q,L}(0) \\ \mathrm{diam}(\mathcal{X}) \sum_{j=l}^{L+1} q^{1-j} \sum_{x' \in Q_j} \mathbb{E}\left[\left|\widehat{\mu}^L_{t,s}(\mathcal{C}(x')) - \mu^L(\mathcal{C}(x'))\right|\right], & \text{if } 2h_{q,L}(l) \leq C < 2h_{q,L}(l-1), \\ \frac{C^p}{2}\mathbb{E}\left[\mathrm{TV}(\widehat{\mu}_{t,s}, \mu)\right], & \text{if } C \leq \left(2h_{q,L}(L) \vee \min_{x \neq x'} d(x, x')\right). \end{cases}$$

For the estimator from (8) with $B_x \sim \mathrm{Ber}(s)$ and $P_x \sim \mathrm{Poi}(t\mu(x))$ for all $x \in \mathcal{X}$, it holds

$$\begin{aligned}
&\sum_{x\in Q_l} \mathbb{E}\left[\left|\widehat{\mu}^L_{t,s}(\mathcal{C}(x)) - \mu^L(\mathcal{C}(x))\right|\right] \\
&= \sum_{x\in Q_l} \frac{1}{st}\mathbb{E}\left[\left|\sum_{y\in\mathcal{C}(x)} P_y B_y - st\sum_{y\in\mathcal{C}(x)}\mu(y)\right|\right] \\
&\leq \sum_{x\in Q_l} \frac{1}{st}\sqrt{\mathrm{Var}\left(\sum_{y\in\mathcal{C}(x)} P_y B_y\right)} = \sum_{x\in Q_l} \frac{1}{st}\sqrt{\sum_{y\in\mathcal{C}(x)}\mathrm{Var}(P_y B_y)} \\
&= \sum_{x\in Q_l} \frac{1}{st}\sqrt{\sum_{y\in\mathcal{C}(x)} s(1-s)t\mu(y) + s(1-s)\mu(y)^2 + t\mu(y)s^2} \\
&= \sum_{x\in Q_l} \sqrt{\frac{1-s}{st}\sum_{y\in\mathcal{C}(x)}\mu(y) + \frac{1-s}{s}\sum_{y\in\mathcal{C}(x)}\mu(y)^2 + \frac{1}{t}\sum_{y\in\mathcal{C}(x)}\mu(y)} \\
&= \sum_{x\in Q_l} \sqrt{\frac{1}{st}\mu\left(\mathcal{C}(x)\right) + \frac{1-s}{s}\sum_{y\in\mathcal{C}(x)}\mu(y)^2} \\
&\leq \sqrt{|Q_l|}\sqrt{\frac{1}{st}\sum_{x\in Q_l}\mu(\mathcal{C}(x)) + \frac{1-s}{s}\sum_{x\in Q_l}\sum_{y\in\mathcal{C}(x)}\mu(y)^2} \\
&= \sqrt{|Q_l|}\sqrt{\frac{1}{st}\mathbb{M}(\mu) + \frac{1-s}{s}\sum_{x\in\mathcal{X}}\mu(x)^2}
\end{aligned}$$

Following an analogous computation one bounds the estimation error for the total mass intensity as

$$\mathbb{E}\left[|\mathbb{M}(\widehat{\mu}_{t,s}) - \mathbb{M}(\mu)|\right] \leq \sqrt{\mathrm{Var}(\mathbb{M}(\widehat{\mu}_{t,s}))} \leq \sqrt{\frac{1}{st}\mathbb{M}(\mu) + \frac{1-s}{s}\sum_{x\in\mathcal{X}}\mu(x)^2}. \tag{31}$$

Applying both of these bounds to the previous upper bound on the (p, C)-KRD in Lemma 1 yields the claim.
For $p > 1$, we first observe again that if $C \leq \min_{x\neq x'} d(x, x')$ then according to (Heinemann et al., 2023, Theorem 2.2) it holds that

$$\mathbb{E}\left[\mathrm{KR}^p_{p,C}(\widehat{\mu}_{t,s}, \mu)\right] = \frac{C^p}{2}\mathbb{E}\left[\mathrm{TV}(\widehat{\mu}_{t,s}, \mu)\right]$$

which yields the total variation bounds. For more general C, we repeat the previous calculations with the upper bounds on the difference of height function $h_{q,L}(j-1)^p - h_{q,L}(j)^p \leq \mathrm{diam}(\mathcal{X})^p \left(\frac{q}{q-1}\right)^p q^{p-jp}$. Since $h_{q,L}(L+1) = 0$ we also have

$$h_{q,L}(L)^p - h_{q,L}(L+1)^p = \mathrm{diam}(\mathcal{X})^p q^{-Lp}.$$

The expectations are bounded identically as before. Finally, using Jensen's inequality,

$$\mathbb{E}\left[\mathrm{KR}_{p,C}(\mu, \widehat{\mu}_{t,s})\right] \leq \left(\mathbb{E}\left[\mathrm{KR}_{p,C}^{p}(\mu, \widehat{\mu}_{t,s})\right]\right)^{\frac{1}{p}},$$

finishes the proof. ■

Remark 35 *Omitting Jensen's inequality in the last step of the proof of Theorem 5 implies the slightly stronger result*

$$\mathbb{E}\left[\mathrm{KR}_{p,C}^{p}\left(\widehat{\mu}_{t,s}, \mu\right)\right] \leq \mathcal{E}_{p,\mathcal{X},\mu}^{\mathrm{Poi}}(C, q, L) \begin{cases} \Big(2(1-s)\mathbb{M}(\mu) + \frac{s}{\sqrt{t}} \sum_{x \in \mathcal{X}} \sqrt{\mu(x)}\Big), \\ \qquad\qquad \text{if } C \leq \big(2h_{q,L}(L) \vee \min_{x \neq x'} d(x, x')\big), \\ \big(\frac{1}{st}\mathbb{M}(\mu) + \frac{1-s}{s} \sum_{x \in \mathcal{X}} \mu(x)^2\big)^{\frac{1}{2}}, \\ \qquad\qquad \text{else.} \end{cases}$$

D.3 Proof of Stability Bound for Balanced Optimal Transport Plan

Proof [Proof of Theorem 14] The proof is divided into several steps. In the first step we cast the optimal transport problem as an appropriate linear program for which we can utilize the Lipschitz stability bound by Li (1994). In the second step we analyze the optimization problem which underlies the definition of respective Lipschitz constant. Finally, in the third step we quantify the Lipschitz constant.

Step 1. Reduction to linear program. The collection of OT plans $\widetilde{\mathbf{P}}_c^*(\mu, \nu)$ for measures μ and ν for cost function c can be interpreted as the solution of the linear program

$$\min_{\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{X})} \sum_{x, x' \in \mathcal{X}} c(x, x')\pi(x, x') \text{ subject to } \begin{cases} \pi(x, x') \geq 0 \text{ for all } x, x' \in \mathcal{X}, \\ \sum_{x' \in \mathcal{X}} \pi(x, x') = \mu(x) \text{ for all } x \in \mathcal{X}, \\ \sum_{x \in \mathcal{X}} \pi(x, x') = \nu(x') \text{ for all } x' \in \mathcal{X}. \end{cases} \tag{32}$$

According to (Luenberger et al., 1984, Theorem, p. 148), one of the summation constraints is redundant and by dropping one of them all remaining one are become linearly independent. We drop the constraint for $\nu(x_{|\mathcal{X}|})$ and the description of the feasible set reduces to

$$\begin{cases} \pi(x, x') \geq 0 \text{ for all } x, x' \in \mathcal{X}, \\ \sum_{x' \in \mathcal{X}} \pi(x, x') = \mu(x) \text{ for all } x \in \mathcal{X}, \\ \sum_{x \in \mathcal{X}} \pi(x, x') = \nu(x') \text{ for all } x' \in \mathcal{X} \backslash \{x_{|\mathcal{X}|}\}. \end{cases}$$

Following the notation of Li (1994), by enumerating the elements of $\mathcal{X}$ as $x_1, \ldots, x_{|\mathcal{X}|}$, we identify π as a vector in $\mathbb{R}^n$ with $n = |\mathcal{X}|^2$ with entries $\pi_{i+|\mathcal{X}|(j-1)} = \pi(x_i, x_j)$ and the marginal measures μ, ν as a vector in $\mathbb{R}^{|\mathcal{X}|}$ with entries $\mu_i = \mu(x_i)$ and $\nu_i = \nu(x_i)$. The

constraints of the optimization problem (32) can be rewritten as $A\pi \leq b$ and $C\pi = d$ for suitable matrices $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{k\times n}$ vectors $b \in \mathbb{R}^n$ and $d \in \mathbb{R}^k$ with $k = 2|\mathcal{X}| - 1$. Herein, the matrix A and the vector b are given by

$$A = -I_n \quad \text{and} \quad b = 0.$$

whereas the matrix C and the vector d are given by

$$C = \begin{pmatrix} I_{|\mathcal{X}|} & I_{|\mathcal{X}|} & \cdots & I_{|\mathcal{X}|} & I_{|\mathcal{X}|} \\ 1_{|\mathcal{X}|} & 0 & \cdots & 0 & 0 \\ 0 & 1_{|\mathcal{X}|} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1_{|\mathcal{X}|} & 0 \end{pmatrix} \quad \text{and} \quad d = d(\mu,\nu) = \begin{pmatrix} \mu(x_1) \\ \vdots \\ \mu(x_{|\mathcal{X}|}) \\ \nu(x_1) \\ \vdots \\ \nu(x_{|\mathcal{X}|-1}) \end{pmatrix}.$$

The objective function in (32) can be written as $c^T\pi$ with $c \in \mathbb{R}^n$ and $c_{i+|\mathcal{X}|(j-1)} = c(x_i, x_j)$. With this notation, the collection of OT plans $\widetilde{\mathbf{P}}_c^*$ is characterized as the solutions of the linear program

$$\min_{\pi\in\mathbb{R}^n} c^T\pi \text{ subject to } A\pi \leq b \text{ and } C\pi = d.$$

Invoking Theorems 2.5 and 3.3 of Li (1994) then yields the stability estimate

$$\begin{aligned} \mathcal{H}_{\mathrm{TV}}\left(\widetilde{\mathbf{P}}^*_{p,C}(\mu^1,\nu^1), \widetilde{\mathbf{P}}^*_{p,C}(\mu^2,\nu^2)\right) &\leq \gamma \|d(\mu^1,\nu^1) - d^2(\mu^2,\nu^2)\| \\ &= \gamma\left(\mathrm{TV}(\mu^1,\mu^2) + \mathrm{TV}(\nu^1,\nu^2)\right), \end{aligned}$$

where $\gamma = \gamma(A, C) > 0$ is defined as

$$\gamma = \sup_{(p,u,v)\in\mathbb{R}^{|\mathcal{X}|^2}\times\mathbb{R}^{|\mathcal{X}|}\times\mathbb{R}^{|\mathcal{X}|-1}} \|(p,u,v)\|_\infty$$
$$\text{subject to} \quad \begin{cases} \|A^Tp + C^T(u^T,v^T)^T\|_\infty \leq 1, \text{ and the rows of } A \\ \text{corresponding to non-zero entries of } p \\ \text{and the rows of } C \text{ are linearly independent.} \end{cases}$$

Hence, the assertion follows once we show that $\gamma \leq 4|\mathcal{X}|$.

Step 2. Analysis of optimization problem for definition of Lipschitz constant. To show the bound, note for $(p,u,v) \in \mathbb{R}^{|\mathcal{X}|^2}\times\mathbb{R}^{|\mathcal{X}|}\times\mathbb{R}^{|\mathcal{X}|-1}$ that $(A^Tp + C^T(u^T,v^T)^T) \in \mathbb{R}^n$ has entries

$$(A^Tp + C^T(u^T,v^T)^T)_{i+|\mathcal{X}|(j-1)} = \begin{cases} -p_{i+|\mathcal{X}|(j-1)} + u_i + v_j, & \text{if } j < |\mathcal{X}|, \\ -p_{i+|\mathcal{X}|(j-1)} + u_i, & \text{if } j = |\mathcal{X}| \end{cases} \quad \text{for } i,j \in |\mathcal{X}|.$$

To guarantee the constraint in the definition of γ, consider some slack variable $z \in \mathbb{R}^n$ with $\|z\|_\infty \leq 1$ and suppose that

$$(A^Tp + C^T(u^T,v^T)^T)_{i+|\mathcal{X}|(j-1)} = z_{i+|\mathcal{X}|(j-1)} \quad \text{for } i,j \in \{1,\ldots,|\mathcal{X}|\}. \tag{33}$$

As is, this system of equation is underdetermined, but by imposing that some entries of p are equal to zero such that the corresponding rows of A and the rows of C are independent, the system either becomes uniquely solvable or unsolvable. The latter case can be ignored, as it does not contribute to the constraint set in the definition of γ.

Since the matrix A has rank $n = |\mathcal{X}|^2$ while C has rank $k = 2|\mathcal{X}| - 1$, it follows that exactly k entries of p need to be set equal to zero for (33) to admit a unique solution.

For k such zero-entries (i, j) of p it follows that (33) reduces to

$$z_{i+|\mathcal{X}|(j-1)} = C^T(u^T, v^T)^T_{i+|\mathcal{X}|(j-1)} = \begin{cases} u_i + v_j, & \text{if } j < |\mathcal{X}|, \\ u_i, & \text{if } j = |\mathcal{X}|. \end{cases} \tag{34}$$

Now, given a pair of solutions for (34), we can recover the remaining entries of p via

$$p_{i+|\mathcal{X}|(j-1)} = -z_{i+|\mathcal{X}|(j-1)} + \begin{cases} u_i + v_j, & \text{if } j < |\mathcal{X}|, \\ u_i, & \text{if } j = |\mathcal{X}| \end{cases}. \tag{35}$$

Hence, for there to be a unique solution (p, u, v) of (33), it is necessary that u and v are uniquely determined by (34). This can only happen if at least one entry $p_{i^*+|\mathcal{X}|(|\mathcal{X}|-1)}$ for $i^* \in \{1, \dots, |\mathcal{X}|\}$ (i.e., for $j = |\mathcal{X}|$) is equal to zero, since otherwise given a solution $(u, v) \in \mathbb{R}^k$ we could construct an alternative solution via $(u + \delta 1_k, v - \delta 1_k) \in \mathbb{R}^k$ for $\delta > 0$.

Step 3. Derivation of upper bound for Lipschitz constant With the previous insight, we infer from (34) that for every $i \in \{1, \dots, |\mathcal{X}|\}$ for which $p_{i+|\mathcal{X}|(|\mathcal{X}|-1)} = 0$ that

$$u_i = z_{i+|\mathcal{X}|(|\mathcal{X}|-1)} \in [-1, 1].$$

Moreover, for some $j \in \{1, \dots, |\mathcal{X}| - 1\}$ such that $p_{i+|\mathcal{X}|(j-1)} = 0$ it follows that

$$v_j = z_{i+|\mathcal{X}|(|\mathcal{X}|-1)} - u_i \in [-2, 2].$$

Since the system of equations (34) admits a unique solution, it follows that every entry u_i and v_j can be obtained by chain of equations. Since there are in total $2|\mathcal{X}| - 1$ equations and from each equation the bound in absolute value for u_i or v_j increases by $|z_{i+|\mathcal{X}|(j-1)}| \leq 1$, it follows altogether that

$$\|(u^T, v^T)^T\|_\infty \leq 2|\mathcal{X}| - 1 \leq 4|\mathcal{X}|.$$

Based on these insights, we infer from (33) for all non-zero entries of p via (35) that

$$|p_{i+|\mathcal{X}|(j-1)}| \leq 4|X| - 2 + 1 \leq 4|\mathcal{X}|.$$

This finally confirms the desired bound for γ and proves the claim. ■

D.4 Proof of Statistical Deviation Bound for Empirical Unbalanced Optimal Transport Plans

Lemma 36 (Relation between restricted and unrestricted UOT plans) *Let $(\mathcal{X}, d)$ be a finite space, $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$. Then, for $p \geq 1$ and $C \geq 0$ the sets $\overline{\mathbf{P}}^*_{p,C}(\mu,\nu)$ defined in (20) and $\mathbf{P}^*_{p,C}(\mu,\nu)$ defined in (21) are related as follows.*

1. *It holds that $\mathbf{P}^*_{p,C}(\mu,\nu) \subseteq \overline{\mathbf{P}}^*_{p,C}(\mu,\nu)$, i.e., the set $\mathbf{P}^*_{p,C}(\mu,\nu)$ consists of unbalanced optimal transport plans.*
2. *Every unbalanced optimal transport plan $\pi \in \overline{\mathbf{P}}^*_{p,C}(\mu,\nu)$ can be represented as the sum of a restricted transport plan $\mathbf{P}^*_{p,C}(\mu,\nu)$ and a non-negative measure supported on $\{(x,x') \in \mathcal{X}\times\mathcal{X} \,|\, d(x,x') = C\}$.*
3. *Every sub-coupling $\pi \in \Pi_{\leq}(\mu,\nu)$ which is the sum of a restricted transport plan $\mathbf{P}^*_{p,C}(\mu,\nu)$ and a non-negative measure supported on $\{(x,x') \in \mathcal{X}\times\mathcal{X} \,|\, d(x,x') = C\}$ is also an unbalanced optimal transport plan, i.e., $\pi \in \overline{\mathbf{P}}^*_{p,C}(\mu,\nu)$.*

Proof For the first claim consider $\pi \in \mathbf{P}^*_{p,C}(\mu,\nu)$ and let $\overline{\pi} \in \overline{\mathbf{P}}^*_{p,C}(\mu,\nu)$ be such that $\pi = \overline{\pi}|_{\mathcal{D}(C)}$. Then it follows that

$$\begin{aligned}
&\sum_{x,x'\in\mathcal{X}} d^p(x,x')\pi(x,x') + C^p\left(\frac{\mathbb{M}(\mu)+\mathbb{M}(\nu)}{2} - \mathbb{M}(\pi)\right)\\
&= \sum_{x,x'\in\mathcal{X}} d^p(x,x')\pi(x,x') + C^p\mathbb{M}(\overline{\pi}-\pi) + C^p\left(\frac{\mathbb{M}(\mu)+\mathbb{M}(\nu)}{2} - \mathbb{M}(\overline{\pi})\right)\\
&= \sum_{x,x'\in\mathcal{X}} d^p(x,x')\overline{\pi}(x,x') + C^p\left(\frac{\mathbb{M}(\mu)+\mathbb{M}(\nu)}{2} - \mathbb{M}(\overline{\pi})\right) = \mathrm{KR}_{p,C}(\mu,\nu),
\end{aligned}$$

where the last equality follows from the fact that $\overline{\pi} \in \overline{\mathbf{P}}^*_{p,C}(\mu,\nu)$. Hence, π is an UOT plan, which shows the first assertion.

For the second claim, we prove that every UOT plan $\overline{\pi} \in \overline{\mathbf{P}}^*_{p,C}(\mu,\nu)$ assigns no mass to the set $\mathcal{E}(C) \coloneqq \{(x,x') \in \mathcal{X}\times\mathcal{X} \mid d(x,x') > C\}$. To this end, consider the decomposition $\overline{\pi} = \overline{\pi}|_{\mathcal{E}(C)} + \overline{\pi}|_{\mathcal{E}(C)^c}$. Then,

$$\begin{aligned}
\mathrm{KR}_{p,C}(\mu,\nu) &= \sum_{x,x'\in\mathcal{X}} d^p(x,x')\overline{\pi}(x,x') + C^p\left(\frac{\mathbb{M}(\mu)+\mathbb{M}(\nu)}{2} - \mathbb{M}(\overline{\pi})\right)\\
&\geq \sum_{(x,x')\in\mathcal{E}(C)} d^p(x,x')\overline{\pi}(x,x') + C^p\mathbb{M}(\overline{\pi}|_{\mathcal{E}(C)^c}) + C^p\left(\frac{\mathbb{M}(\mu)+\mathbb{M}(\nu)}{2} - \mathbb{M}(\overline{\pi})\right)\\
&= \sum_{x,x'\in\mathcal{X}} d^p(x,x')\overline{\pi}|_{\mathcal{E}(C)}(x,x') + C^p\left(\frac{\mathbb{M}(\mu)+\mathbb{M}(\nu)}{2} - \mathbb{M}(\overline{\pi}|_{\mathcal{E}(C)})\right) \geq \mathrm{KR}_{p,C}(\mu,\nu).
\end{aligned}$$

This implies that $\overline{\pi}|_{\mathcal{E}(C)^c} = 0$, which shows the second assertion.

Finally, for the third claim, consider $\pi \in \overline{\mathbf{P}}^*_{p,C}(\mu,\nu)$ and $\widetilde{\pi} \in \mathcal{M}_+(\mathcal{X})$ which is supported on $\{(x,x') \in \mathcal{X} \times \mathcal{X} \mid d(x,x') = C\}$, and assume that $\overline{\pi} \coloneqq \pi + \widetilde{\pi} \in \Pi_{\leq}(\mu,\nu)$. Then,

$$\begin{aligned}
\mathrm{KR}_{p,C}(\mu,\nu) &\leq \sum_{x,x' \in \mathcal{X}} d^p(x,x')\overline{\pi}(x,x') + C^p \left(\frac{\mathbb{M}(\mu) + \mathbb{M}(\nu)}{2} - \mathbb{M}(\overline{\pi}) \right) \\
&= \sum_{x,x' \in \mathcal{X}} d^p(x,x')\pi(x,x') + C^p\widetilde{\pi}(x,x') + C^p \left(\frac{\mathbb{M}(\mu) + \mathbb{M}(\nu)}{2} - \mathbb{M}(\pi) - \mathbb{M}(\widetilde{\pi}) \right) \\
&= \sum_{x,x' \in \mathcal{X}} d^p(x,x')\pi(x,x') + C^p \left(\frac{\mathbb{M}(\mu) + \mathbb{M}(\nu)}{2} - \mathbb{M}(\pi) \right) = \mathrm{KR}_{p,C}(\mu,\nu),
\end{aligned}$$

which asserts optimality of $\overline{\pi}$. ■

Lemma 37 *Let $(\mathcal{X},d)$ be a finite metric space and let $p \geq 1$ and $C \geq 0$. Then, the following assertions hold.*

1. *For every pair of measures $\mu,\nu \in \mathcal{M}_+(\mathcal{X})$ and their augmented counterparts $\widetilde{\mu},\widetilde{\nu} \in \mathcal{M}_+^B(\widetilde{\mathcal{X}}) = \{\mu \in \mathcal{M}_+(\mathcal{X}) \mid \mathbb{M}(\mu) \leq B\}$ for $B = \max(\mathbb{M}(\mu),\mathbb{M}(\nu))$ it holds that*

$$\mathbf{P}^*_{p,C}(\mu,\nu) = \left\{ \widetilde{\pi} \cdot \mathbb{1}(\cdot \in \mathcal{D}(C)) \,\middle|\, \widetilde{\pi} \in \widetilde{\mathbf{P}}^*_{p,C}(\widetilde{\mu},\widetilde{\nu}) \right\}. \tag{36}$$

2. *For every two pairs of measures $\mu^1,\nu^1,\mu^2,\nu^2 \in \mathcal{M}_+(\mathcal{X})$ and corresponding augmented measures $\widetilde{\mu}^i,\widetilde{\nu}^i \in \mathcal{M}_+^{B_i}(\widetilde{\mathcal{X}})$ for $B_i = \max(\mathbb{M}(\mu^i),\mathbb{M}(\nu^i))$ and $i \in \{1,2\}$ it holds that*

$$\mathcal{H}_{\mathrm{TV}}\left(\mathbf{P}^*_{p,C}(\mu^1,\nu^1), \mathbf{P}^*_{p,C}(\mu^2,\nu^2)\right) \leq \mathcal{H}_{\mathrm{TV}}\left(\widetilde{\mathbf{P}}^*_{p,C}(\widetilde{\mu}^1,\widetilde{\nu}^1), \widetilde{\mathbf{P}}^*_{p,C}(\widetilde{\mu}^2,\widetilde{\nu}^2)\right).$$

Proof For the first assertion assume without loss of generality that $\mathbb{M}(\mu) \geq \mathbb{M}(\nu)$ which implies $\widetilde{\mu} = \mu$. Further, throughout the proof we will make use of the following equality which has been shown by Heinemann et al. (2023, p. 17) and is discussed in Appendix C of this manuscript,

$$\mathrm{KR}^p_{p,C}(\mu,\nu) = \mathrm{UOT}_{d^p \wedge C^p, C}(\mu,\nu) = \widetilde{\mathrm{OT}}_{\widetilde{d}^p_C}(\widetilde{\mu},\widetilde{\nu}). \tag{37}$$

To show that the right-hand side of (36) is included in the left-hand side, let $\widetilde{\pi} \in \widetilde{\mathbf{P}}^*_{p,C}(\widetilde{\mu},\widetilde{\nu})$ and define $\pi \coloneqq \widetilde{\pi} \cdot \mathbb{1}(\cdot \in \mathcal{D}(C))$ and $\pi^* \coloneqq \widetilde{\pi} \cdot \mathbb{1}(\cdot \in \mathcal{X} \times \mathcal{X})$. Note that $\widetilde{\pi}(\{\mathfrak{d}\} \times \widetilde{\mathcal{X}}) = \widetilde{\mu}(\{\mathfrak{d}\}) = 0$. As a result, by the marginal constraints on $\widetilde{\pi}$, $\widetilde{\pi}(\widetilde{\mathcal{X}} \times \mathcal{X}) = \widetilde{\nu}(\mathcal{X}) = \nu(\mathcal{X})$, it

follows that $\mathbb{M}(\pi^*) = \mathbb{M}(\nu)$. From this it is evident that

$$\begin{aligned}
&\sum_{x,x'\in\mathcal{X}} d^p(x,x')\pi(x,x') + C^p\left(\frac{\mathbb{M}(\mu)+\mathbb{M}(\nu)}{2} - \mathbb{M}(\pi)\right)\\
&= \sum_{x,x'\in\mathcal{X}} d^p(x,x')\pi(x,x') + C^p\mathbb{M}(\pi^*-\pi) + C^p\left(\frac{\mathbb{M}(\mu)+\mathbb{M}(\nu)}{2} - \mathbb{M}(\pi^*)\right)\\
&= \sum_{x,x'\in\mathcal{X}} \widetilde{d}_C^p(x,x')\pi^*(x,x') + \frac{C^p}{2}\left(\mathbb{M}(\mu)-\mathbb{M}(\nu)\right)\\
&= \sum_{x,x'\in\mathcal{X}} \widetilde{d}_C^p(x,x')\widetilde{\pi}(x,x') + \sum_{x\in\mathcal{X}} \widetilde{d}_C^p(x,\mathfrak{d})\widetilde{\pi}(x,\mathfrak{d})\\
&= \sum_{x,x'\in\widetilde{\mathcal{X}}} \widetilde{d}_C^p(x,x')\widetilde{\pi}(x,x') = \widetilde{\mathrm{OT}}_{\widetilde{d}_C^p}(\widetilde{\mu},\widetilde{\nu}) = \mathrm{KR}_{p,C}^p(\mu,\nu),
\end{aligned}$$

where the equality third to last is a consequence of $\widetilde{\pi}(\mathfrak{d}\times\widetilde{\mathcal{X}}) = \widetilde{\mu}(\mathfrak{d}) = 0$, the second to last equality follows by optimality of $\widetilde{\pi}$, and the final equality is a consequence of (37). In conjunction with Lemma 36 this asserts that $\pi \in \mathbf{P}_{p,C}^*(\mu,\nu)$.

Conversely, to show that the left-hand side of (36) is contained in the right-hand side, let $\pi \in \mathbf{P}_{p,C}^*(\mu,\nu)$. Further, define the measure $\check{\pi} \in \mathcal{M}_+(\mathcal{X}\times\mathcal{X})$ for $x,x' \in \mathcal{X}$ by

$$\check{\pi}(x,x') \coloneqq \begin{cases} 0, & \text{if } \mathbb{M}(\nu) = \mathbb{M}(\pi),\\ \frac{\mu(x)-\pi(x,\mathcal{X})}{\mathbb{M}(\mu)-\mathbb{M}(\pi)}\left(\nu(x') - \pi(\mathcal{X},x')\right) & , \text{ if } \mathbb{M}(\nu) > \mathbb{M}(\pi), \end{cases}$$

and further set $\pi^* \coloneqq \pi + \check{\pi}$. Then, it follows that $\pi^* \in \Pi_{\leq}(\mu,\nu)$ and $\mathbb{M}(\pi^*) = \mathbb{M}(\nu)$. Further, from optimality of π we infer that

$$\begin{aligned}
\mathrm{KR}_{p,C}^p(\mu,\nu) &= \sum_{x,x'\in\mathcal{X}} d^p(x,x')\pi(x,x') + C^p\left(\frac{\mathbb{M}(\mu)+\mathbb{M}(\nu)}{2} - \mathbb{M}(\pi)\right)\\
&= \sum_{x,x'\in\mathcal{X}} (d^p(x,x')\wedge C^p)\pi(x,x') + C^p\mathbb{M}(\check{\pi}) + C^p\left(\frac{\mathbb{M}(\mu)+\mathbb{M}(\nu)}{2} - \mathbb{M}(\pi+\check{\pi})\right)\\
&\geq \sum_{x,x'\in\mathcal{X}} (d^p(x,x')\wedge C^p)\pi^*(x,x') + C^p\left(\frac{\mathbb{M}(\mu)+\mathbb{M}(\nu)}{2} - \mathbb{M}(\pi^*)\right)\\
&\geq \mathrm{UOT}_{d^p\wedge C^p,C}(\mu,\nu) = \mathrm{KR}_{p,C}^p(\mu,\nu), \qquad (38)
\end{aligned}$$

where the last equality follows (37) and thus asserts that all inequalities are qualities. This implies that $\check{\pi}$ is concentrated on $\{(x,x') \in \mathcal{X}\times\mathcal{X} \mid d(x,x') \geq C\}$ and, thus, $\pi^* \cdot \mathbb{1}(\cdot \in \mathcal{D}(C)) = \pi$. Next, define $\widetilde{\pi} \coloneqq \pi^* + \sum_{x\in\mathcal{X}} \delta_{(x,\mathfrak{d})}\left(\mu(x) - \pi^*(x,\mathcal{X})\right)$ which fulfills $\widetilde{\pi} \in \Pi_{=}(\widetilde{\mu},\widetilde{\nu})$ and $\widetilde{\pi}\cdot\mathbb{1}(\cdot\in\mathcal{D}(C)) = \pi$. In particular, from the above derivation in (38) we infer from

$\mathbb{M}(\pi^*) = \mathbb{M}(\nu)$, $\mathbb{M}(\widetilde{\pi}) = \mathbb{M}(\widetilde{\mu}) = \mathbb{M}(\mu)$ and since $\widetilde{\pi}(\mathfrak{d}, \widetilde{\mathcal{X}}) = \widetilde{\mu}(\mathfrak{d}) = 0$ that

$$\begin{aligned}
\mathrm{KR}^p_{p,C}(\mu,\nu) &= \sum_{x,x' \in \mathcal{X}} (d^p(x,x') \wedge C^p)\pi^*(x,x') + C^p \left(\frac{\mathbb{M}(\mu) + \mathbb{M}(\nu)}{2} - \mathbb{M}(\pi^*) \right) \\
&= \sum_{x,x' \in \mathcal{X}} (d^p(x,x') \wedge C^p)\pi^*(x,x') + \frac{C^p}{2} \left(\mathbb{M}(\mu) - \mathbb{M}(\nu) \right) \\
&= \sum_{x,x' \in \mathcal{X}'} \widetilde{d}^p_C(x,x')\pi^*(x,x') \geq \widetilde{\mathrm{OT}}_{\widetilde{d}^p_C}(\widetilde{\mu},\widetilde{\nu}) = \mathrm{KR}^p_{p,C}(\mu,\nu).
\end{aligned}$$

Here, the final equality is a consequence of (37), and altogether confirms that $\widetilde{\pi} \in \widetilde{\mathbf{P}}^*_{p,C}(\widetilde{\mu},\widetilde{\nu})$. This concludes the proof of the first assertion.

For the proof of the second assertion note from the considerations above that for every pair of UOT plans $\pi^i \in \mathbf{P}^*(\mu^i,\nu^i)$ with $i \in \{1,2\}$, every pair of corresponding OT plans $\widetilde{\pi}^i \in \widetilde{\mathbf{P}}^*(\mu^i,\nu^i)$ with $\widetilde{\pi}^i \cdot \mathbb{1}(\cdot \in \mathcal{D}(C)) = \pi^i$ fulfills the inequality

$$\mathrm{TV}(\pi^1,\pi^2) \leq \mathrm{TV}(\widetilde{\pi}^1,\widetilde{\pi}^2).$$

Based on this, the claim follows from the definition of the Hausdorff distance induced by the total variation norm. ■

Proof [Proof of Theorem 11] For the proof we again rely on the lift to optimal transport as detailed in Appendix C. To this end, consider the augmented space $\widetilde{\mathcal{X}} = \mathcal{X} \cup \{\mathfrak{d}\}$ for some dummy element $\mathfrak{d}$, let B be a majorant for the masses of the measures $\mu^1, \mu^2, \nu^1, \nu^2$, and define for $i \in \{1,2\}$ the augmented measures

$$\begin{aligned}
\widetilde{\mu}^i &:= \mu^i + \delta_{\mathfrak{d}}(\max(\mathbb{M}(\mu^i), \mathbb{M}(\nu^i)) - \mathbb{M}(\mu^i)) \quad \text{and} \\
\widetilde{\nu}^i &:= \nu^i + \delta_{\mathfrak{d}}(\max(\mathbb{M}(\mu^i), \mathbb{M}(\nu^i)) - \mathbb{M}(\nu^i)).
\end{aligned}$$

Then, upon denoting the collection of OT plans between $\widetilde{\mu}^i$ and $\widetilde{\nu}^i$ with respect to the cost function $\widetilde{d}^p_C$ by $\widetilde{\mathbf{P}}^*_{p,C}(\widetilde{\mu}^i,\widetilde{\nu}^i)$, it follows by Lemma 37 and Theorem 14 that

$$\begin{aligned}
\mathcal{H}_{\mathrm{TV}}\left(\mathbf{P}^*_{p,C}(\mu^1,\nu^1), \mathbf{P}^*_{p,C}(\mu^2,\nu^2) \right) &\leq \mathcal{H}_{\mathrm{TV}}\left(\widetilde{\mathbf{P}}^*_{p,C}(\widetilde{\mu}^1,\widetilde{\nu}^1), \widetilde{\mathbf{P}}^*_{p,C}(\widetilde{\mu}^2,\widetilde{\nu}^2) \right) \\
&\leq 4\,|\widetilde{\mathcal{X}}| \left(\mathrm{TV}(\widetilde{\mu}^1,\widetilde{\mu}^2) + \mathrm{TV}(\widetilde{\nu}^1,\widetilde{\nu}^2) \right).
\end{aligned}$$

The assertion now follows by $|\widetilde{\mathcal{X}}| = |\mathcal{X}| + 1$ and the subsequent Lemma 38. ■

Lemma 38 *Let $\mu^i, \nu^i \in \mathcal{M}_+(\mathcal{X})$ for $i \in \{1,2\}$ be two non-negative measures and consider their augmented measures $\widetilde{\mu}^i, \widetilde{\nu}^i \in \mathcal{M}^B_+(\widetilde{\mathcal{X}})$ to $\widetilde{\mathcal{X}}$ for $B = \max(\mathbb{M}(\mu^i), \mathbb{M}(\nu^i))$, then*

$$\begin{aligned}
&\mathrm{TV}(\widetilde{\mu}^1,\widetilde{\mu}^2) + \mathrm{TV}(\widetilde{\nu}^1,\widetilde{\nu}^2) \\
&\leq \mathrm{TV}(\mu^1,\mu^2) + \mathrm{TV}(\nu^1,\nu^2) + |\mathbb{M}(\mu^1) - \mathbb{M}(\mu^2)| + |\mathbb{M}(\nu^1) - \mathbb{M}(\nu^2)| \\
&\leq 2\left(\mathrm{TV}(\mu^1,\mu^2) + \mathrm{TV}(\nu^1,\nu^2) \right).
\end{aligned}$$

Proof We first observe that

$$\begin{aligned}
&\mathrm{TV}(\widetilde{\mu}^1, \widetilde{\mu}^2) + \mathrm{TV}(\widetilde{\nu}^1, \widetilde{\nu}^2) \\
&= \left(\sum_{x \in \mathcal{X}} |\mu^1(x) - \mu^2(x)| + |\nu^1(x) - \nu^2(x)| \right) + |\widetilde{\mu}^1(\mathfrak{d}) - \widetilde{\mu}^2(\mathfrak{d})| + |\widetilde{\nu}^1(\mathfrak{d}) - \widetilde{\nu}^2(\mathfrak{d})| \\
&= \mathrm{TV}(\mu^1, \mu^2) + \mathrm{TV}(\nu^1, \nu^2) + |\widetilde{\mu}^1(\mathfrak{d}) - \widetilde{\mu}^2(\mathfrak{d})| + |\widetilde{\nu}^1(\mathfrak{d}) - \widetilde{\nu}^2(\mathfrak{d})|.
\end{aligned}$$

Now if $\mathbb{M}(\mu^1) \geq \mathbb{M}(\nu^1)$ and $\mathbb{M}(\mu^2) \geq \mathbb{M}(\nu^2)$, then

$$\begin{aligned}
|\widetilde{\mu}^1(\mathfrak{d}) - \widetilde{\mu}^2(\mathfrak{d})| + |\widetilde{\nu}^1(\mathfrak{d}) - \widetilde{\nu}^2(\mathfrak{d})| &= |\mathbb{M}(\mu^1) - \mathbb{M}(\nu^1) - \mathbb{M}(\mu^2) + \mathbb{M}(\nu^2)| \\
&\leq |\mathbb{M}(\mu^1) - \mathbb{M}(\mu^2)| + |\mathbb{M}(\nu^1) - \mathbb{M}(\nu^2)|,
\end{aligned}$$

whereas if $\mathbb{M}(\mu^1) \geq \mathbb{M}(\nu^1)$ and $\mathbb{M}(\mu^2) < \mathbb{M}(\nu^2)$, then

$$\begin{aligned}
|\widetilde{\mu}^1(\mathfrak{d}) - \widetilde{\mu}^2(\mathfrak{d})| + |\widetilde{\nu}^1(\mathfrak{d}) - \widetilde{\nu}^2(\mathfrak{d})| &= |\mathbb{M}(\nu^1) - \mathbb{M}(\mu^1)| + |\mathbb{M}(\mu^2) - \mathbb{M}(\nu^2)| \\
&= \mathbb{M}(\nu^1) - \mathbb{M}(\mu^1) + \mathbb{M}(\mu^2) - \mathbb{M}(\nu^2) \\
&\leq |\mathbb{M}(\mu^1) - \mathbb{M}(\mu^2)| + |\mathbb{M}(\nu^1) - \mathbb{M}(\nu^2)|.
\end{aligned}$$

The remaining two cases can be treated analogously, asserting the first inequality of the claim. For the second inequality it suffices to note that the total variation norm provides an upper bound on the mass difference of measures ■

D.5 Proof of Statistical Deviation Bound for Empirical $(\mathbf{p}, \mathbf{C})$-Barycenters

Proof [Proof of Theorem 16, Theorem 26 and Theorem 33] Let $\widehat{\mu}^1, \ldots, \widehat{\mu}^J$ be any of the three estimators from (6), (7) or (8). Further, for each $1 \leq i \leq J$ let $\mathcal{E}_{1,\mathcal{X}_i,\mu^i}(C)$ be the corresponding constant in Theorem 5, Theorem 22 or Theorem 30 for μ^i, respectively. Let θ_i denote the respective sampling parameter dependencies $N_i^{-1/2}$, $\phi(t_i, s_i)$ or $\psi(s_{\mathcal{X}_i})$, respectively. Involving the augmentation argument, due to the construction of the lifted problem, it holds for any $\mu^1, \mu^2, \mu^3 \in \mathcal{M}_+(\mathcal{Y})$ that

$$\begin{aligned}
|\mathrm{KR}^p_{p,C}(\mu^1, \mu^3) - \mathrm{KR}^p_{p,C}(\mu^2, \mu^3)| &= |\widetilde{\mathrm{OT}}^p_p(\widetilde{\mu}^1, \widetilde{\mu}^3) - \widetilde{\mathrm{OT}}^p_p(\widetilde{\mu}^2, \widetilde{\mu}^3)| \\
&\leq \mathrm{diam}(\widetilde{\mathcal{Y}})^{p-1} p \, \widetilde{\mathrm{OT}}_1(\widetilde{\mu}^1, \widetilde{\mu}^2) \\
&= C^{p-1} p \, \mathrm{KR}_{1,C}(\mu^1, \mu^2),
\end{aligned}$$

where the inequality follows from Sommerfeld and Munk (2018). Taking expectation and applying the previous display together with Theorem 5 yields

$$\begin{aligned}
\mathbb{E}\left[|F_{p,C}(\mu) - \widehat{F}_{p,C}(\mu)| \right] &\leq \frac{1}{J} \sum_{i=1}^{J} \mathbb{E}\left[|\mathrm{KR}^p_{p,C}(\mu^i, \mu) - \mathrm{KR}^p_{p,C}(\widehat{\mu}^i, \mu)| \right] \\
&\leq p \, C^{p-1} \frac{1}{J} \sum_{i=1}^{J} \mathbb{E}\left[\mathrm{KR}_{1,C}(\mu^i, \widehat{\mu}^i) \right] \\
&\leq p C^{p-1} \frac{1}{J} \sum_{i=1}^{J} \mathcal{E}_{1,\mathcal{X}_i,\mu^i}(C) \theta_i.
\end{aligned}$$

Let $\mu^\star$ and $\widehat{\mu}^\star$ be minimizers of their respective p-Fréchet functional $F_{p,C}$ and $\widehat{F}_{p,C}$. Then, it follows that

$$\begin{aligned}
\mathbb{E}[|F_{p,C}(\widehat{\mu}^\star) - F_{p,C}(\mu^\star)|] &= \mathbb{E}\left[F_{p,C}(\widehat{\mu}^\star) - \widehat{F}_{p,C}(\mu^\star) + \widehat{F}_{p,C}(\mu^\star) - F_{p,C}(\mu^\star)\right] \\
&\leq \mathbb{E}\left[F_{p,C}(\widehat{\mu}^\star) - \widehat{F}_{p,C}(\mu^\star)\right] + \mathbb{E}\left[\widehat{F}_{p,C}(\mu^\star) - F_{p,C}(\mu^\star)\right] \\
&\leq \mathbb{E}\left[F_{p,C}(\widehat{\mu}^\star) - \widehat{F}_{p,C}(\mu^\star)\right] + p\, C^{p-1} \frac{1}{J} \sum_{i=1}^{J} \mathcal{E}_{1,\mathcal{X}_i,\mu^i}(C)\theta_i \\
&\leq \mathbb{E}\left[F_{p,C}(\widehat{\mu}^\star) - \widehat{F}_{p,C}(\widehat{\mu}^\star)\right] + p\, C^{p-1} \frac{1}{J} \sum_{i=1}^{J} \mathcal{E}_{1,\mathcal{X}_i,\mu^i}(C)\theta_i \\
&\leq p\, C^{p-1} \frac{2}{J} \sum_{i=1}^{J} \mathcal{E}_{1,\mathcal{X}_i,\mu^i}(C)\theta_i,
\end{aligned}$$

where the fourth inequality follows from $\widehat{\mu}^*$ being a minimizer of $\widehat{F}_{p,C}$. ■

Proof [Proof of Theorem 17, Theorem 27 and Theorem 34] Let $\widehat{\mu}^1, \dots, \widehat{\mu}^J$, $\mathcal{E}_{1,\mathcal{X}_i,\mu^i}(C)$ and θ_i for all $i = 1, \dots, J$ as in the previous proof. Let $\mathbf{B}$ be the set of (p, C)-barycenters of the measures $\mu^1, \dots, \mu^J$ and define $\widetilde{\mathbf{B}}$ as the set of OT_p-barycenters of the augmented measures $\widetilde{\mu}^1, \dots, \widetilde{\mu}^J$. Similar, we denote $\widehat{\mathbf{B}}$ the set of (p, C)-barycenters of the estimated measures $\widehat{\mu}^1, \dots, \widehat{\mu}^J$ and let $\widehat{\widetilde{\mathbf{B}}}$ be the set of p-barycenters of their augmented versions. Define the lift of a measure $\mu \in \mathcal{M}_+(\mathcal{Y})$ to a measure $\widetilde{\mu} \in \mathcal{M}(\widetilde{\mathcal{Y}})$ by

$$\phi_{\mu^1,\dots,\mu^J}(\mu) = \mu + \left(\sum_{i=1}^{J} \mathbb{M}(\mu^i) - \mathbb{M}(\mu)\right) \delta_{\mathfrak{d}}.$$

If $\mu \in \mathbf{B}$ then it follows by (Heinemann et al., 2023, Lemma 3.3) that $\phi_{\mu^1,\dots,\mu^J}(\mu) \in \widetilde{\mathbf{B}}$. Conversely, for any $\widetilde{\mu} \in \widetilde{\mathbf{B}}$ it holds that $\phi^{-1}_{\mu^1,\dots,\mu^J}(\widetilde{\mu}) \in \mathbf{B}$. We denote by $\phi(\mathbf{B}) \coloneqq \{\phi_{\mu^1,\dots,\mu^J}(\mu) | \mu \in \mathbf{B}\}$ and analogously $\phi^{-1}\left(\widetilde{\mathbf{B}}\right) \coloneqq \{\phi^{-1}_{\mu^1,\dots,\mu^J}(\widetilde{\mu}) | \widetilde{\mu} \in \widetilde{\mathbf{B}}\}$. With this we have

$$\begin{aligned}
\mathbb{E}\left[\sup_{\widehat{\mu} \in \widehat{\mathbf{B}}} \inf_{\mu \in \mathbf{B}} \mathrm{KR}^p_{p,C}(\mu, \widehat{\mu})\right] &= \mathbb{E}\left[\sup_{\widehat{\mu} \in \phi^{-1}\left(\widehat{\widetilde{\mathbf{B}}}\right)} \inf_{\mu \in \phi^{-1}\left(\widetilde{\mathbf{B}}\right)} \mathrm{KR}^p_{p,C}(\mu, \widehat{\mu})\right] && (39) \\
&= \mathbb{E}\left[\sup_{\widehat{\mu} \in \phi^{-1}\left(\widehat{\widetilde{\mathbf{B}}}\right)} \inf_{\mu \in \phi^{-1}\left(\widetilde{\mathbf{B}}\right)} \widetilde{\mathrm{OT}}^p_p(\phi(\mu), \phi(\widehat{\mu}))\right] && (40) \\
&= \mathbb{E}\left[\sup_{\widehat{\mu} \in \widehat{\widetilde{\mathbf{B}}}} \inf_{\mu \in \widetilde{\mathbf{B}}} \widetilde{\mathrm{OT}}^p_p(\widetilde{\mu}, \widehat{\widetilde{\mu}})\right].
\end{aligned}$$

We continue by recalling a slightly adapted version of Lemma 3.8 in Heinemann et al. (2022). Since we only apply this lemma to the augmented balanced OT problem, the proof remains unchanged and is therefore omitted.

Lemma 39 *Let $\widetilde{F}_{p,C}$ be the augmented Fréchet functional corresponding to $\widetilde{\mu}^1, \dots, \widetilde{\mu}_N \in \mathcal{M}_+(\widetilde{\mathcal{Y}})$. Then, for any $\widetilde{\mu} \in M_+(\mathcal{C}_{\mathrm{KR}}(J, p, C) \cup \{\mathfrak{d}\})$ with $\mathbb{M}(\widetilde{\mu}) = \sum_{i=1}^J \mathbb{M}(\mu^i)$ there exists a $\widetilde{\mu}^* \in \operatorname{argmin}_{\widetilde{\nu}} \widetilde{F}_{p,C}(\widetilde{\nu})$ such that*

$$\widetilde{F}_{p,C}(\widetilde{\mu}) - \widetilde{F}_{p,C}(\widetilde{\mu}^*) \geq 2V_P \widetilde{\mathrm{OT}}_p^p(\widetilde{\mu}, \widetilde{\mu}^*),$$

where V_P is the constant from Theorem 17.

Invoking Lemma 39 with $\widehat{\mu} \in \widehat{\widehat{\mathbf{B}}}$ and applying Theorem 16 yields

$$\begin{aligned}\frac{2p}{J} \sum_{i=1}^J \mathcal{E}_{1,\mathcal{X}_i,\mu^i}(C) C^{p-1} \theta &\geq \mathbb{E}\left[F_{p,C}(\widehat{\widetilde{\mu}}) - F_{p,C}(\widetilde{\mu})\right] \\ &\geq \mathbb{E}\left[2V_p \sup_{\widehat{\widetilde{\mu}} \in \widehat{\widetilde{\mathbf{B}}}} \inf_{\widetilde{\mu} \in \widetilde{\mathbf{B}}} \widetilde{\mathrm{OT}}_p^p\left(\widetilde{\mu}, \widehat{\widetilde{\mu}}\right)\right] = \mathbb{E}\left[2V_p \sup_{\widehat{\mu} \in \widehat{\mathbf{B}}} \inf_{\mu \in \mathbf{B}} \mathrm{KR}_{p,C}^p(\mu, \widehat{\mu})\right],\end{aligned}$$

where the equality follows from (39) and hence

$$\frac{\frac{p}{J} \sum_{i=1}^J \mathcal{E}_{1,\mathcal{X}_i,\mu^i}(C)\ C^{p-1}}{V_P} \theta \geq \mathbb{E}\left[\sup_{\widehat{\mu} \in \widehat{\mathbf{B}}} \inf_{\mu \in \mathbf{B}} \mathrm{KR}_{p,C}^p(\mu, \widehat{\mu})\right].$$

■

Appendix E. Additional Simulations

In this appendix we display additional simulations which showcase the relative error in estimating the $(2, C)$-KRD and the $(2, C)$-KR barycenter for the Poisson model as well as for the multinomial and Bernoulli model.

E.1 Additional Synthetic Datasets

This subsection introduces four additional classes of measures SPI, SPIC, and NI, NIG. The first two are qualitatively similar to the classes NE and NEC, the latter two resemble weighted modifications of the classes PI and PIC.

Spirals of varying Length (SPI), Figure 17 (a)

Let $a_i \sim U[2, 4]$ and $b_i \sim U[3, 6]$ for $i = 1, \dots, J$. Let $K_i = \lceil b_i M \rceil$ and let $t_1, \dots, t_K$ be a discretization of $[0, b\pi]$. Set $w_k^i = 1$ for $k = 1, \dots, K_i$ and $i = 1, \dots, J$. Set

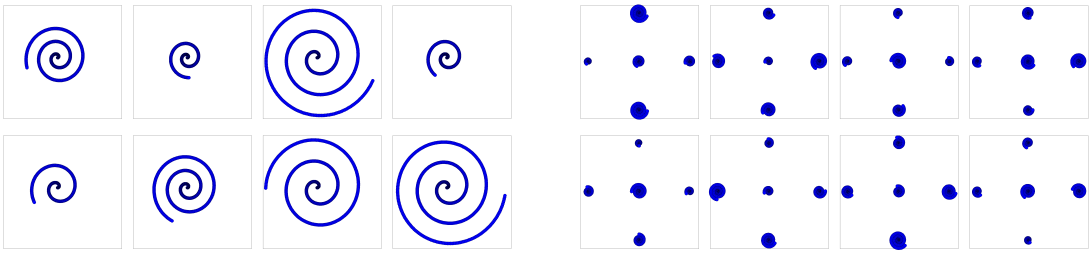
$$l_k^i = a_i((t_k \sin(t_k) + 64)/140, (t_k \cos(t_k) + 70)/130)^T.$$

Clustered Spirals (SPIC), see Figure 17 (b)

Let $a_i^c \sim U[2,4]$ and $b_i^c \sim U[3,6]$ for $i = 1,\dots,J$, $c = 1,\dots,5$. Let $K_i^c = \lceil b_i^c M \rceil$ and let $t_1,\dots,t_K$ be a discretization of $[0, b\pi]$. Set $w_k^i = 1$ for $k = 1,\dots,K_i$ and $i = 1,\dots,J$ and let $\alpha = (0,3,3,6,3)^T$ and $\beta = (3,0,3,3,6)^T$. Set

$$l^i_{\sum_{r=1}^{c-1} K_i^r + k} = (1/7)(((a_i^c t_k \sin(t_k) + 64)/140) + \alpha_c, (a_i^c t_k \cos(t_k) + 70)/130 + \beta_c)^T,$$

where we use the convention that a sum is zero if its last index is smaller than its first one.



(a) (b)

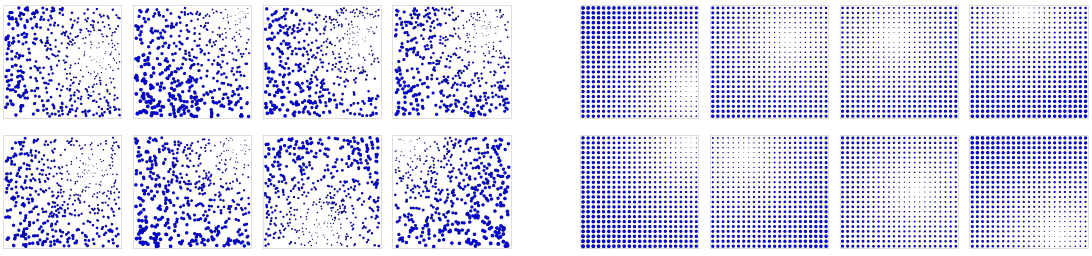
Figure 17: **(a)** An example of $J = 8$ measures from the *SPI* dataset with $M = 110$. **(b)** An example of $J = 8$ measures from the *SPIC* dataset with $M = 22$.

Norm-based Intensities on Uniform Positions (NI), see Figure 18 (a)

Fix J locations $l_0^1,\dots,l_0^J \in [0,1]^2$. Let $l_1^i,\dots,l_K^i \sim U[0,1]^2$ and let $w_k^i = \|l_k^i - l_0^i\|_2$ for $1 \le i \le J$.

Norm-based Intensities on a Grid (NIG), see Figure 18 (b)

Let $K = M^2$ for $M \in \mathbb{N}$. Fix J locations $l_0^1,\dots,l_0^J \in [0,1]^2$ and let $l_1^i,\dots,l_{M^2}^i$ be the points of an equidistant $M \times M$ grid in $[0,1]^2$ for each $1 \le i \le J$. Set $w_k^i = \|l_k^i - l_0^i\|_2$ for $1 \le i \le J$.



(a) (b)

Figure 18: **(a)** An example of $J = 8$ measures from the *NI* dataset with $M = 500$. **(b)** An example of $J = 8$ measures from the *NIG* dataset with $M = 23^2$.

E.2 Simulations for the Poisson Model

This section details the simulation results for the additional classes of measures SPI, SPIC, NI, and NIG. Figures 19 – 22 display the relative error of the empirical KRD in estimating the population KRD across several choices for $C \in \{0.01, 0.1, 1, 10\}$. Moreover, Figures 23 – 26 detail the relative Fréchet error in estimating the $(2, C)$-KR barycenter for the choices $C \in \{0.1, 1, 10\}$.

Structurally, the simulation results and corresponding conclusions for the SPI and SPIC classes are similar to those for the respective NE and NEC classes. Likewise, the insights about the NI and NIG classes are closely tied to those of the PI and PIG classes, respectively.

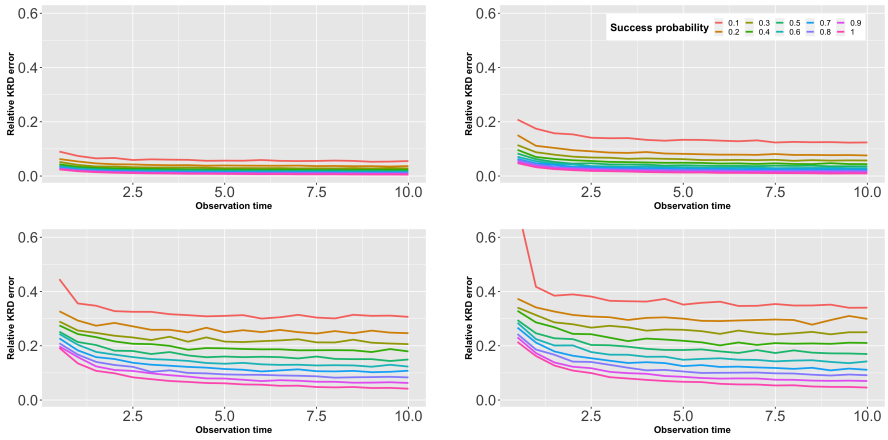


Figure 19: As in Figure 10, but for the SPI class and $M = 65$.

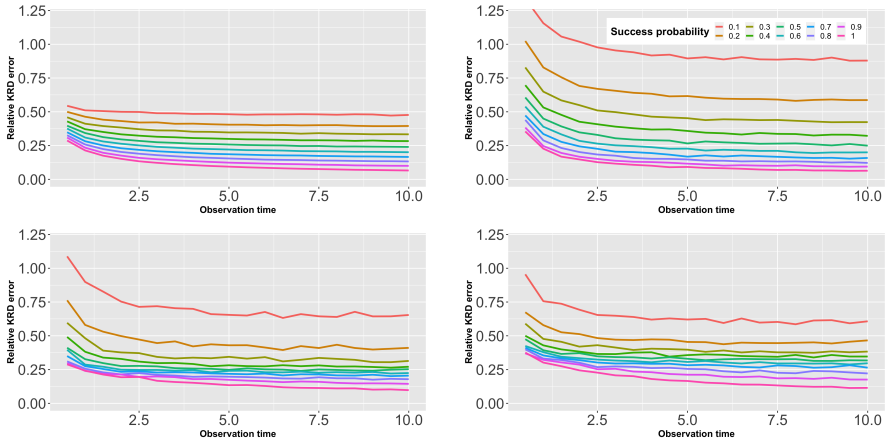


Figure 20: As in Figure 10, but for the SPIC class and $M = 12$.

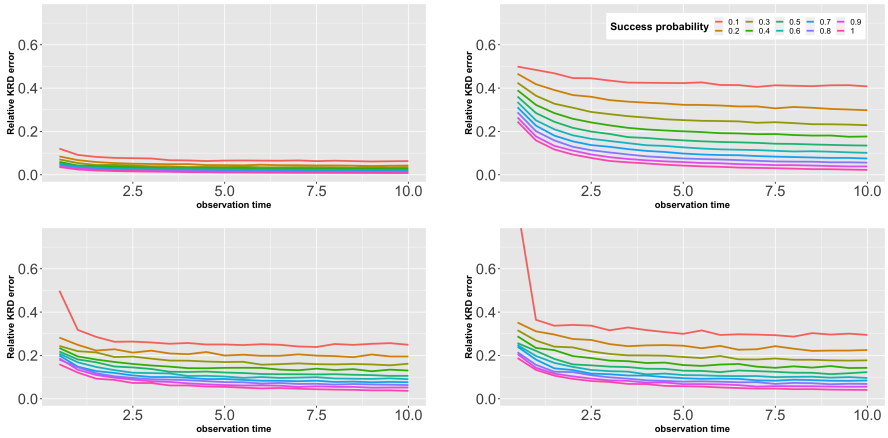


Figure 21: As in Figure 10, but for the NI class and $M = 300$.

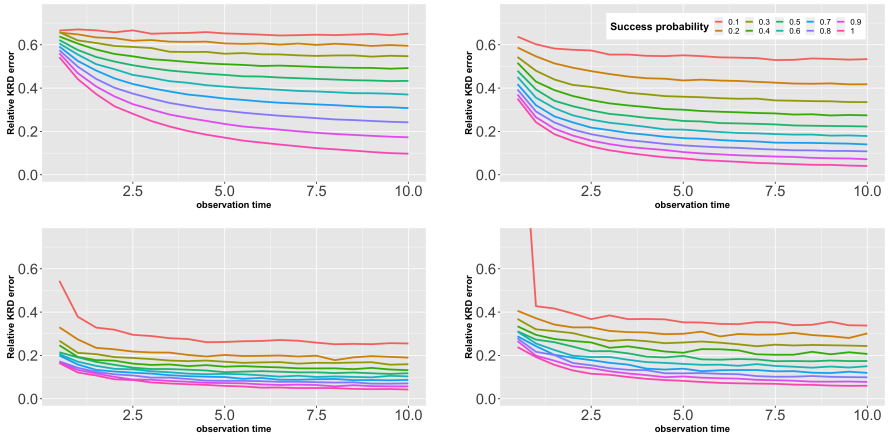


Figure 22: As in Figure 10, but for the NIG class and $M = 17$.

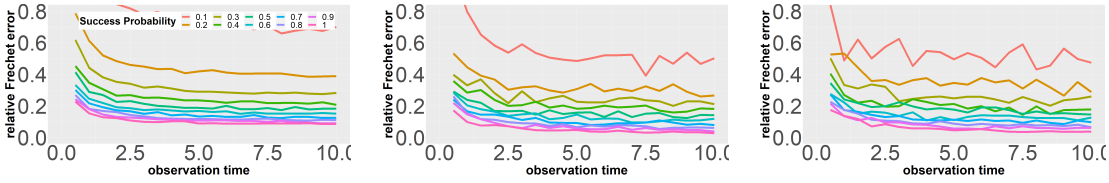


Figure 23: As in Figure 15, but for the SPI class and $M = 65$.

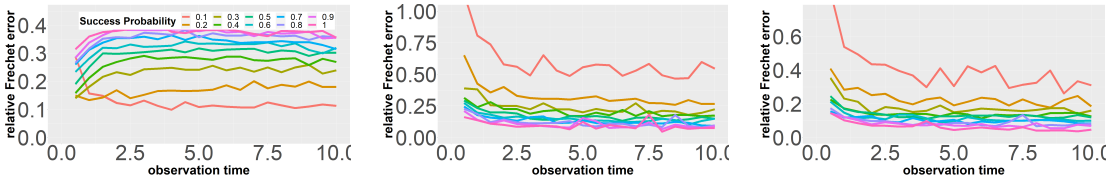


Figure 24: As in Figure 15, but for the SPIC class and $M = 12$.

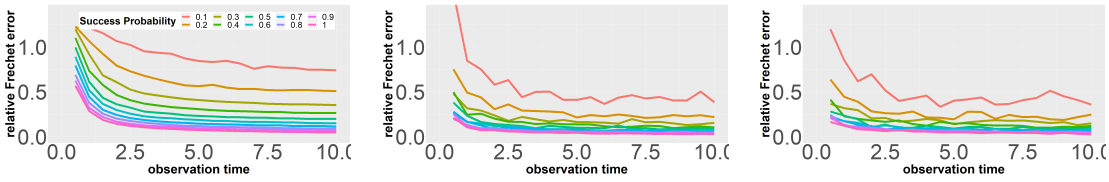


Figure 25: As in Figure 15, but for the NI class and $M = 300$.

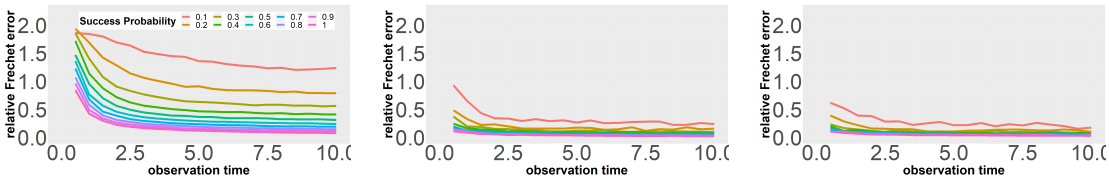


Figure 26: As in Figure 15, but for the NIG class and $M = 17$.

E.3 Simulations for Multinomial Model

In this subsection, we repeat the simulations from Section 6 and Appendix E.2 for the multinomial model. We observe that with increasing sample size N, the statistical error mostly decreases with rate $N^{-1/2}$ which is what our theory in Appendix A asserts. For the (p, C)-KRD the results (in Figure 27) slightly differ from the results in the Poisson model. Notably, for increasing C, the error is decreasing. This is explained by the fact that in the multinomial scheme, we do not have to estimate the total intensities of the measures, and it is precisely this estimation error that drives the error for increasing C in the Poisson model. Similarly to the Poisson scheme, we observe a decrease in error for the measure classes with clustered support structures when C surpasses the distance between two individual clusters. For the (p, C)-barycenters under the multinomial sampling model (in Figure 28) there is an initial increase in error for small sample sizes. Specifically, this occurs for $C = 0.1$ and the NEC and SPIC classes. This value of C is below the cluster size. This effect is most likely for these measure classes. For increasing C there is a significant reduction in estimation error. In particular, for some classes the error reduces by two orders of magnitude going from $C = 0.1$ to $C = 10$. Since the total mass intensities of the individual measures do not need to be estimated in this sampling model, we already observe a decrease in error for increasing C for the (p, C)-KRD and naturally there is a similar effect for the Fréchet functional.

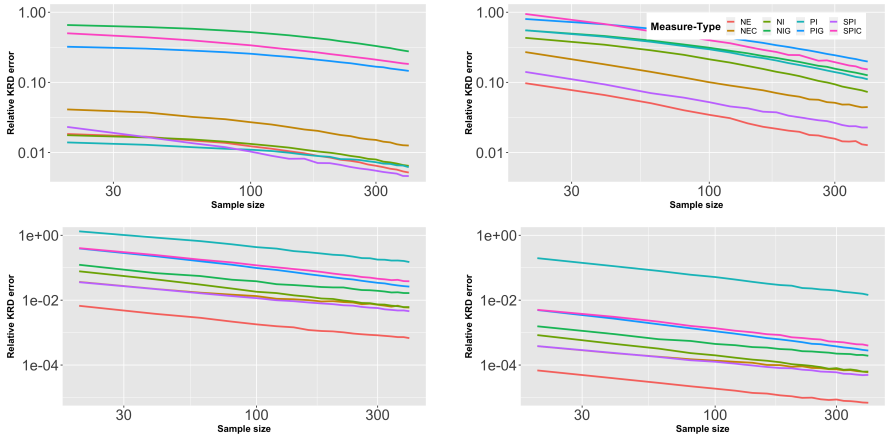


Figure 27: Log-log-plots of expected relative $(2, C)$-KRD error for two measures in the multinomial model for the classes in Section 6. For each sampling size N the expectation is estimated from 1000 independent runs. For each class the parameters are set, such that the measures have on average 300 support points. From top-left to bottom-right we have $C = 0.01, 0.1, 1, 10$, respectively.

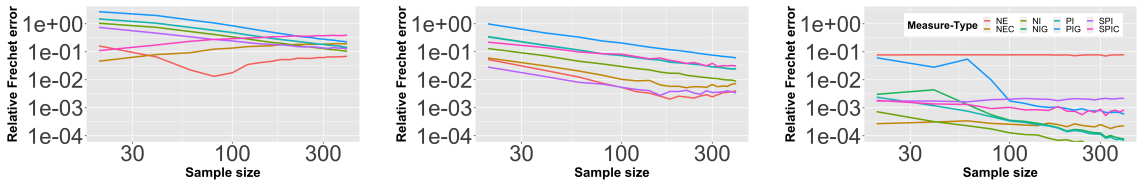


Figure 28: Log-log-plots of expected relative Fréchet error for the $(2, C)$-barycenter for $J = 5$ measures from the PI class for the multinomial model with different sample sizes N. For each sample size the expectation is estimated from 100 independent runs. For each class the parameters are set, such that the measures have on average 300 support points. From left to right we have $C = 0.1, 1, 10$, respectively.

E.4 Simulations for Bernoulli Model

To construct a reasonable framework for the simulations in the Bernoulli model, define for $s_0 \in \mathbb{R}_+$,

$$s_x = \frac{s_0}{\|x - (0.5, 0.5)^T\|_2 + s_0}.$$

Intuitively, the success probability at a given point x is larger, if x is closer to the center of $[0,1]^2$ and smaller if it is further away from the center. Further, for $s_0 \to \infty$ the success probability at each location converges to one. For the simulations, we now consider the error as a function of s_0. Note, that in this simulation study only the classes of measures with mass one at each support point are considered in accordance with the Bernoulli model in (7). One notable observation for the empirical (p, C)-KRD (in Figure 29) is that the error of the SPIC class is significantly higher than for the NEC class, even though they share the same cluster locations. This can be explained by the fact that, by construction, the measures in the NEC class have a higher proportion of their mass in their central clusters, which is close to $(0.5, 0.5)^T$ and thus has a high probability of being observed. This effect also carries over to the (p, C)-barycenter (in Figure 30). In general, for the (p, C)-KRD the error in this model is increasing in C (which is again explained by the estimation error for the true total mass intensity). However, the effect is less pronounced than in the Poisson model. For the clustered data types a small decrease of error for increasing C over the cluster size can again be noted. Though, also this effect is less significant than in the other models. For the (p, C)-barycenter a decrease in error in C can be observed which is consistent with the previous results for the Poisson model and again explained by the increased stability of the total mass intensity of the barycenter compared to the individual ones.

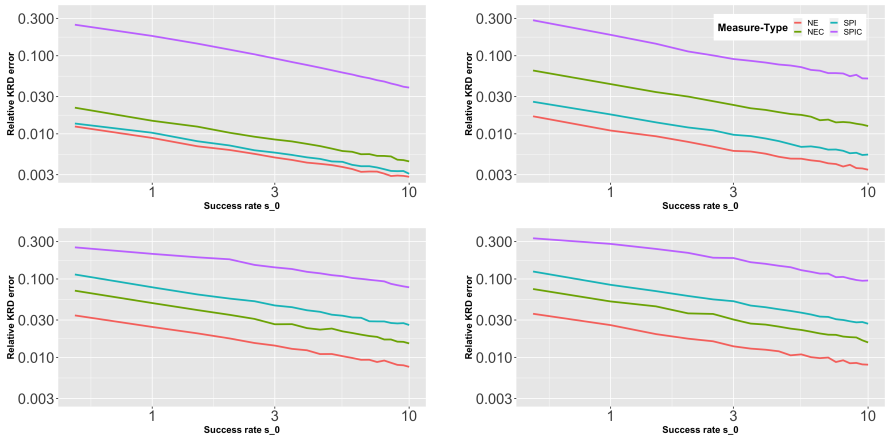


Figure 29: Log-log-plots of expected relative $(2, C)$-KRD error in terms of success probability s_0 for two measures in the Bernoulli Model for the NE, NEC, SPI, and SPIC classes from Section 6 and Appendix E.1. For each success probability s_0 the expectation is estimated from 1000 independent runs. The parameters are chosen such that the population measures have on average 300 support points. From top-left to bottom-right we have $C = 0.01, 0.1, 1, 10$, respectively.

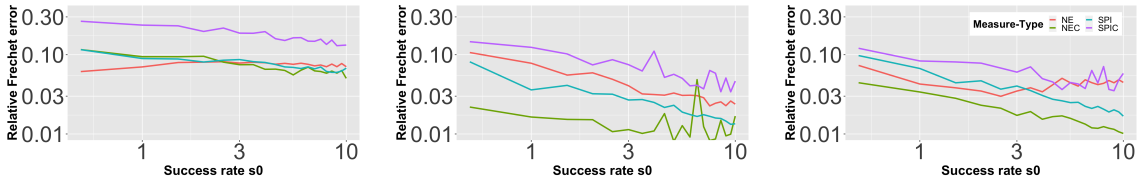


Figure 30: Expected relative Fréchet error for the $(2, C)$-barycenter in terms of success probability s_0 for $J = 5$ measures in the Bernoulli Model for the NE, NEC, SPI, and SPIC classes from Section 6 and Appendix E.1. For each success probability s_0 the expectation is estimated from 100 independent runs. The parameters are set such that the population measures in all classes have on average 300 support points. From left to right we have $C = 0.1, 1, 10$, respectively.

References

M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

J. Altschuler, J. Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1961–1971, 2017.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. Proceedings of Machine Learning Research, 2017.

T. Aspelmeier, A. Egner, and A. Munk. Modern statistical challenges in high-resolution fluorescence microscopy. *Annual Review of Statistics and Its Application*, 2:163–202, 2015.

Y. Balaji, R. Chellappa, and S. Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 33:12934–12944, 2020.

M. Bansil and J. Kitagawa. Quantitative stability in the geometry of semi-discrete optimal transport. *International Mathematics Research Notices*, 2022(10):7354–7389, 2022.

C. F. Baumgartner, L. M. Koch, K. C. Tezcan, J. X. Ang, and E. Konukoglu. Visual feature attribution using Wasserstein GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8309–8319, 2018.

D. Berend and A. Kontorovich. A sharp estimate of the binomial mean absolute deviation with applications. *Statistics & Probability Letters*, 83(4):1254–1259, 2013.

D. Bertsimas and J. N. Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

E. Boissard and T. Le Gouic. On the mean speed of convergence of empirical and occupation measures in wasserstein distance. *Annales de l'IHP Probabilités et statistiques*, 50(2):539–563, 2014.

N. Bonneel, M. Van De Panne, S. Paris, and W. Heidrich. Displacement interpolation using Lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*, pages 1–12, 2011.

N. Bonneel, G. Peyré, and M. Cuturi. Wasserstein barycentric coordinates: Histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4):71–1, 2016.

C. Bunne, S. G. Stark, G. Gut, J. S. Del Castillo, M. Levesque, K.-V. Lehmann, L. Pelkmans, A. Krause, and G. Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023.

L. Chapel, M. Z. Alaya, and G. Gasso. Partial optimal tranport with applications on positive-unlabeled learning. In H. Larochelle, M. Ranzato, et al., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2903–2913. Curran Associates, Inc., 2020.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018a.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018b.

L. Chizat, P. Roussillon, F. Léger, F.-X. Vialard, and G. Peyré. Faster Wasserstein distance estimation with the Sinkhorn divergence. In H. Larochelle, M. Ranzato, et al., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2257–2269. Curran Associates, Inc., 2020.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26:2292–2300, 2013.

N. Deb, P. Ghosal, and B. Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29736–29753. Curran Associates, Inc., 2021a.

N. Deb, P. Ghosal, and B. Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *Advances in Neural Information Processing Systems*, 34:29736–29753, 2021b.

E. del Barrio, A. González Sanz, and J.-M. Loubes. Central limit theorems for semi-discrete wasserstein distances. *Bernoulli*, 30(1):554–580, 2024.

A. Delalande and Q. Merigot. Quantitative stability of optimal transport maps under variations of the target measure. *Duke Mathematical Journal*, 172(17):3321–3357, 2023.

S. Dereich, M. Scheutzow, and R. Schottstedt. Constructive quantization: Approximation by empirical measures. *Annales de l'IHP Probabilités et statistiques*, 49(4):1183–1203, 2013.

V. Divol, J. Niles-Weed, and A.-A. Pooladian. Tight stability bounds for entropic Brenier maps. *Preprint arXiv:2404.02855*, 2024.

R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *International conference on machine learning*, pages 1367–1376. PMLR, 2018.

S. N. Evans and F. A. Matsen. The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):569–592, 2012.

K. Fatras, Y. Zine, S. Majewski, R. Flamary, R. Gribonval, and N. Courty. Minibatch optimal transport distances; analysis and applications. *Preprint arXiv:2101.01792*, 2021.

A. Figalli. The optimal partial transport problem. *Archive for Rational Mechanics and Analysis*, 195(2):533–560, 2010.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

G. Friesecke, D. Matthes, and B. Schmitzer. Barycenters for the Hellinger–Kantorovich distance over $\mathbb{R}^d$. *SIAM Journal on Mathematical Analysis*, 53(1):62–110, 2021.

C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a Wasserstein loss. *Advances in Neural Information Processing Systems*, 28:2053–2061, 2015.

M. Gellert, M. F. Hossain, F. J. F. Berens, L. W. Bruhn, C. Urbainsky, V. Liebscher, and C. H. Lillig. Substrate specificity of thioredoxins and glutaredoxins – towards a functional classification. *Heliyon*, 5(12):e02943, 2019.

N. Gigli. On Hölder continuity-in-time of the optimal transport map towards measures along a curve. *Proceedings of the Edinburgh Mathematical Society*, 54(2):401–409, 2011.

A. Gramfort, G. Peyré, and M. Cuturi. Fast optimal transport averaging of neuroimaging data. In *International Conference on Information Processing in Medical Imaging*, pages 261–272. Springer, 2015.

K. Guittet. Extended Kantorovich norms: a tool for optimization. Technical report, Technical Report 4402, INRIA, 2002.

W. Guo, N. Ho, and M. Jordan. Fast algorithms for computational optimal transport and Wasserstein barycenter. In *International Conference on Artificial Intelligence and Statistics*, pages 2088–2097. Proceedings of Machine Learning Research, 2020.

M. Hallin. Measure transportation and statistical decision theory. *Annual Review of Statistics and Its Application*, 9:401–424, 2022.

M. Hallin, E. Del Barrio, J. Cuesta-Albertos, and C. Matrán. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139–1165, 2021.

L. G. Hanin. Kantorovich-Rubinstein norm and its application in the theory of Lipschitz spaces. *Proceedings of the American Mathematical Society*, 115(2):345–352, 1992.

F. Heinemann, A. Munk, and Y. Zemel. Randomized Wasserstein barycenter computation: Resampling with statistical guarantees. *SIAM Journal on Mathematics of Data Science*, 4(1):229–259, 2022.

F. Heinemann, M. Klatt, and A. Munk. Kantorovich–rubinstein distance and barycenter for finitely supported measures: Foundations and algorithms. *Applied Mathematics & Optimization*, 87(1):4, 2023.

S. Hundrieser, B. Eltzner, and S. F. Huckemann. A lower bound for estimating fr\'echet means. *Preprint arXiv:2402.12290*, 2024a.

S. Hundrieser, M. Klatt, T. Staudt, and A. Munk. A unifying approach to distributional limits for empirical optimal transport. *Bernoulli*, 30(4):2846–2877, 2024b.

S. Hundrieser, T. Staudt, and A. Munk. Empirical optimal transport between different measures adapts to lower complexity. *Annales de l'IHP Probabilités et statistiques*, 60(2): 824–846, 2024c.

J.-C. Hütter and P. Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194, 2021.

L. V. Kantorovich and S. Rubinstein. On a space of totally additive functions. *Vestnik of the Saint Petersburg University: Mathematics*, 13(7):52–59, 1958.

M. Klatt, C. Tameling, and A. Munk. Empirical regularized optimal transport: Statistical theory and applications. *SIAM Journal on Mathematics of Data Science*, 2(2):419–443, 2020.

M. Klatt, A. Munk, and Y. Zemel. Limit laws for empirical optimal solutions in random linear programs. *Annals of Operations Research*, 315(1):251–278, 2022.

N. Kolkin, J. Salavon, and G. Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10051–10060, 2019.

P. T. Komiske, E. M. Metodiev, and J. Thaler. Metric space of collider events. *Physical review letters*, 123(4):041801, 2019.

G. Kulaitis, A. Munk, and F. Werner. What is resolution? A statistical minimax testing perspective on superresolution microscopy. *The Annals of Statistics*, 49(4):2292–2312, 2021.

K. Le, H. Nguyen, Q. M. Nguyen, T. Pham, H. Bui, and N. Ho. On robust optimal transport: Computational complexity and barycenter computation. In M. Ranzato, A. Beygelzimer, et al., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 21947–21959. Curran Associates, Inc., 2021.

K. Le, H. Nguyen, K. Nguyen, T. Pham, and N. Ho. On multimarginal partial optimal transport: Equivalent forms and computational complexity. In G. Camps-Valls, F. J. R. Ruiz, et al., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 4397–4413. PMLR, 28–30 Mar 2022.

J. Lee and M. Raginsky. Minimax statistical learning with Wasserstein distances. *Advances in Neural Information Processing Systems*, 31, 2018.

W. Li. Sharp Lipschitz constants for basic optimal solutions and basic feasible solutions of linear programs. *SIAM Journal on Control and Optimization*, 32(1):140–153, 1994.

M. Liero, A. Mielke, and G. Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones Mathematicae*, 211(3):969–1117, 2018.

T. Lin, N. Ho, X. Chen, M. Cuturi, and M. Jordan. Fixed-support Wasserstein barycenters: Computational hardness and fast algorithm. *Advances in Neural Information Processing Systems*, 33, 2020.

S. Liu, F. Bunea, and J. Niles-Weed. Asymptotic confidence sets for random linear programs. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3919–3940. PMLR, 2023.

D. G. Luenberger, Y. Ye, et al. *Linear and Nonlinear Programming*, volume 2. Springer, 1984.

T. Manole and J. Niles-Weed. Sharp convergence rates for empirical optimal transport with smooth costs. *The Annals of Applied Probability*, 34(1B):1108–1135, 2024.

T. Manole, S. Balakrishnan, J. Niles-Weed, and L. Wasserman. Plugin estimation of smooth optimal transport maps. *The Annals of Statistics*, 52(3):966–998, 2024a.

T. Manole, P. Bryant, J. Alison, M. Kuusela, and L. Wasserman. Background modeling for double higgs boson production: Density ratios and optimal transport. *The Annals of Applied Statistics*, 18(4):2950–2978, 2024b.

G. Mena and J. Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in Neural Information Processing Systems*, 32, 2019.

D. Mukherjee, A. Guha, J. M. Solomon, Y. Sun, and M. Yurochkin. Outlier-robust optimal transport. In *International Conference on Machine Learning*, pages 7850–7860. Proceedings of Machine Learning Research, 2021.

A. Munk, T. Staudt, and F. Werner. Statistical foundations of nanoscale photonic imaging. In *Nanoscale Photonic Imaging*, pages 125–143. Springer, Cham, 2020.

J. Naas, G. Nies, H. Li, S. Stoldt, B. Schmitzer, S. Jakobs, and A. Munk. Multimatch: Geometry-informed colocalization in multi-color super-resolution microscopy. *bioRxiv*, 2024.

K. Nguyen, D. Nguyen, Q. Nguyen, T. Pham, H. Bui, D. Phung, T. Le, and N. Ho. On transportation of mini-batches: a hierarchical approach. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning 2022*, volume 162, pages 16622–16655, 2022.

J. Niles-Weed and Q. Berthet. Minimax estimation of smooth densities in Wasserstein distance. *The Annals of Statistics*, 50(3):1519–1540, 2022.

V. M. Panaretos and Y. Zemel. Statistical aspects of Wasserstein distances. *Annual review of statistics and its application*, 6:405–431, 2019.

O. Pele and M. Werman. A linear time histogram metric for improved SIFT matching. In *European Conference on Computer Vision*, pages 495–508. Springer, 2008.

O. Pele and M. Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th international conference on computer vision*, pages 460–467. IEEE, 2009.

G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

F. Pitié, A. C. Kokaram, and R. Dahyot. Automated colour grading using colour distribution transfer. *Computer Vision and Image Understanding*, 107(1-2):123–137, 2007.

J. Rabin and N. Papadakis. Convex color image segmentation with optimal transport distances. In *International conference on scale space and variational methods in computer vision*, pages 256–269. Springer, 2015.

Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pages 59–66. IEEE, 1998.

R. Sadhu, Z. Goldfeld, and K. Kato. Limit theorems for semidiscrete optimal transport maps. *The Annals of Applied Probability*, 34(6):5694–5736, 2024.

F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 55 of *Progress in Nonlinear Differential Equations and Their Applications*. Springer, 2015.

G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

M. A. Schmitz, M. Heitz, N. Bonneel, F. Ngole, D. Coeurjolly, M. Cuturi, G. Peyré, and J.-L. Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

S. Singh and B. Póczos. Minimax distribution estimation in Wasserstein distance. *Preprint arXiv:1802.08855*, 2018.

M. Sommerfeld and A. Munk. Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society Series B*, 80(1):219–238, 2018.

M. Sommerfeld, J. Schrieber, Y. Zemel, and A. Munk. Optimal transport: Fast probabilistic approximation with exact solvers. *Journal of Machine Learning Research*, 20(105):1–23, 2019.

T. Staudt and S. Hundrieser. Convergence of empirical optimal transport in unbounded settings. *Bernoulli [To appear, preprint arXiv:2306.11499]*, 2023.

C. Tameling, S. Stoldt, T. Stephan, J. Naas, S. Jakobs, and A. Munk. Colocalization for super-resolution microscopy via optimal transport. *Nature Computational Science*, 1(3): 199–211, 2021.

G. Tartavel, G. Peyré, and Y. Gousseau. Wasserstein loss for image synthesis and restoration. *SIAM Journal on Imaging Sciences*, 9(4):1726–1755, 2016.

A. Vacher, B. Muzellec, A. Rudi, F. Bach, and F.-X. Vialard. A dimension-free computational upper-bound for smooth optimal transport estimation. In *Conference on Learning Theory*, pages 4143–4173, 2021.

E. Ventre, A. Forrow, N. Gadhiwala, P. Chakraborty, O. Angel, and G. Schiebinger. Trajectory inference for a branching SDE model of cell differentiation. *Preprint arXiv:2307.07687*, 2023.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

C. Villani. *Topics in Optimal Transportation*. American Mathematical Soc., 2003.

C. Villani. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

J. Weed. An explicit analysis of the entropic penalty in linear programming. In *Conference On Learning Theory*, pages 1841–1855. PMLR, 2018.

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang. Low-dose ct image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. *IEEE transactions on Medical Imaging*, 37(6): 1348–1357, 2018.